

# The persistent cosmic web and its filamentary structure I: Theory and implementation

T. Sousbie<sup>1,2</sup>

<sup>1</sup>*Department of Physics, The University of Tokyo, Tokyo 113-0033, Japan,*

<sup>2</sup>*Institut d'astrophysique de Paris & UPMC (UMR 7095), 98, bis boulevard Arago, 75 014, Paris.*

*tsousbie@gmail.com, sousbie@utap.phys.s.u-tokyo.ac.jp*

22 September 2010

## ABSTRACT

We present DisPerSE, a novel approach to the coherent multi-scale identification of all types of astrophysical structures, and in particular the filaments, in the large scale distribution of matter in the Universe. This method and corresponding piece of software allows a genuinely scale free and parameter free identification of the voids, walls, filaments, clusters and their configuration within the cosmic web, directly from the discrete distribution of particles in N-body simulations or galaxies in sparse observational catalogues. To achieve that goal, the method works directly over the Delaunay tessellation of the discrete sample and uses the DTFE density computed at each tracer particle; no further sampling, smoothing or processing of the density field is required.

The idea is based on recent advances in distinct sub-domains of computational topology, namely the *discrete* Morse theory which allows a rigorous application of topological principles to astrophysical data sets, and the theory of persistence, which allows us to consistently account for the intrinsic uncertainty and Poisson noise within data sets. Practically, the user can define a given persistence level in terms of robustness with respect to noise (defined as a “number of sigmas”) and the algorithm returns the structures with the corresponding significance as sets of critical points, lines, surfaces and volumes corresponding to the clusters, filaments, walls and voids; filaments, connected at cluster nodes, crawling along the edges of walls bounding the voids. From a geometrical point of view, the method is also interesting as it allows for a robust quantification of the topological properties of a discrete distribution in terms of Betti numbers or Euler characteristics, without having to resort to smoothing or having to define a particular scale.

In this paper, we introduce the necessary mathematical background and describe the method and implementation, while we address the application to 3D simulated and observed data sets to the companion paper, Sousbie, Pichon, Kawahara (2010).

**Key words:** Cosmology: simulations, statistics, observations, Galaxies: formation, dynamics.

## 1 INTRODUCTION

The existence of an intricate network of filaments in the large scale distribution of matter is now considered an established fact. It was first observed by de Lapparent et al. (1986) (see also *e.g.* Colless et al. 2003) and latter theorized (see *e.g.* Pogosyan et al. 1996; Bond et al. 1996): under-dense void regions bounded by sheet-like walls embedded in a web like filamentary network branching on high density dark matter haloes and galaxy clusters form the so called cosmic web Bond et al. (1996), that spans over a wide range of scales larger than the Megaparsec. Dark matter halos and galaxy clusters have arguably been the

most studied component, and there exist a wide range of methods to identify them in simulations or observational catalogues such as the classical friend-of-friend (FOF) (Huchra & Geller 1982), HFOF and 6D minimal spanning tree (Gottloeber 1998), SUBFIND (Springel et al. 2001), VOBOS (Neyrinck et al. 2005) or ADAPTAHOP (Aubert et al. 2004; Tweed et al. 2009) (the list is not exhaustive). Cosmological voids were first observed by Kirshner et al. (1981) and theoretical models were latter developed (see *e.g.* Hoffman & Shaham 1982; Icke 1984; Bertschinger 1985). Although they have been the subject of less attention, there still exist a large number of references describing their features and introducing numerical void finders such

as for instance Neyrinck (2008), Platen et al. (2007) or Aragon-Calvo et al. (2010) (see also the references therein). Because of the intrinsic difficulty of even defining the concepts of wall and filament, not to mention designing consistent identification algorithms (especially in the case of observational data), their generic properties still remain relatively uncertain. One can for instance refer to Aragon-Calvo et al. (2010) for a nice review of the different identification techniques and a study of the filaments properties in dark matter N-body simulations (see also *e.g.* Gay et al. 2010), and Stoica et al. (2010) or Sousbie et al. (2008) for recent attempts at identifying filaments properties in the SDSS and 2dFGRS galaxy catalogues, using the CANDY model (Stoica et al. 2005) and skeleton formalism (Sousbie et al. 2008) respectively. In this paper, we present a general framework within which the physically meaningful objects that are the voids, walls, filaments and haloes are rigorously and consistently defined and we also detail the corresponding numerical method that allows for their direct identification in simulated as well as observational data sets. We focus in particular on what is probably the most striking feature of matter distribution on large scales in the Universe, its filamentary structure.

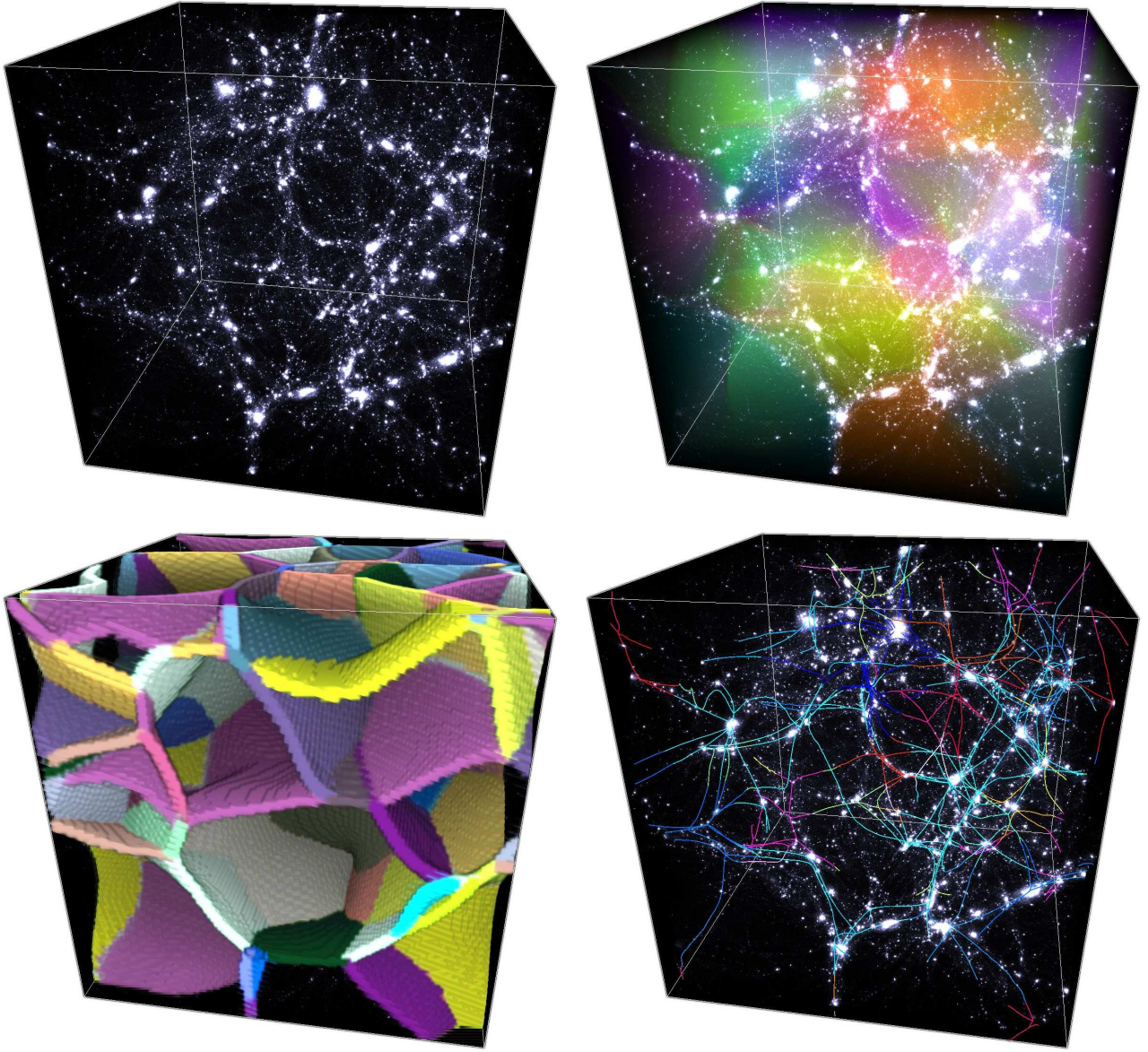
During the last few years, Morse theory (*e.g.* Milnor (1963); Jost (2008)) has been recognized as a very promising approach to the global identification of all types of astrophysically significant features of the large scale galaxy distribution in the universe (see *e.g.* Novikov et al. 2006; Hahn et al. 2007; Sousbie et al. 2008, 2009, 2008; Aragon-Calvo et al. 2008; Forero-Romero et al. 2009). The main reason for this strong interest comes from the fact that all the salient features of the web-like pattern of galaxies have a direct, mathematically well defined equivalent in Morse theory. In fact, Morse theory mainly relies on the definition of so-called ascending and descending  $k$ -manifolds, which *partition* space into series of  $k$ -dimensional domains defined by the gradient of a function (in the present case, the density field), and the network whose branches are formed by their intersections and whose nodes are the critical points, the so-called Morse-complex (see section 2). As illustrated on figure 1, each of those can be directly associated to an astrophysical objects of interest: an ascending 3-manifold defines a void, an ascending 2-manifold defines a wall, and ascending 1-manifold defines a filaments, a descending 3-manifold defines a peak-patch of peak theory (Bardeen et al. 1986), ... and the Morse complex defines some sort of hierarchy and a notion of neighbourhood between them (see section 2 for more details).

Nevertheless, and as promising as it may seem, all the efforts toward applying Morse theory to astrophysical data sets such as galaxy catalogues have so far been plagued by major difficulties. Those difficulties are a direct consequence of the fact that Morse theory, although very attractive, is fundamentally a mathematical theory defined for idealized, well defined and properly behaved smooth functions, which of course is not generally the case of any physical data set resulting from actual measurements. At least two critical issues can be identified in the case of the large scale structure identification problem. The first results from the presence of Poisson noise and large

observational biases in galaxy catalogues, which should be dealt with from the start, especially when the data set is relatively sparse as it becomes even more difficult in that case to distinguish between noise features and the actual features of the sampled data set. The second issue arises from the fact that Morse theory applies to so called Morse functions (see definition 2.2), which are sufficiently smooth twice differentiable *continuous* functions (whose critical points are non-degenerate) whereas the galaxy distribution is discrete by nature. This incompatibility is fundamental, as it means that the theoretical notions of Morse theory may actually not apply to any practical data set. A more detailed discussion of this problem is presented in appendix A as well as an example of the consequences of neglecting this inconsistency in the case of watershed based methods such as Sousbie et al. (2009); Aragon-Calvo et al. (2008).

In this paper, we focus on presenting DisPerSE, a formalism and corresponding software specifically designed for analyzing the cosmic web and its filamentary network. This formalism is based on Morse theory, while the aforementioned incompatibilities with astrophysical data sets are overcome by relying on relatively recent advances in distinct sub-domains of computational topology. These domains are discrete Morse theory (a distinct though related theory developed by Forman see Forman (1998b, 2002) and references therein) and persistent homology, first introduced in Edelsbrunner et al. (2000, 2002). We therefore start by introducing the corresponding necessary notions of computational topology in sections 2, 3 and 4 respectively. Note that no previous knowledge in the field of computational topology is assumed here, the goal of those sections being mainly to introduce the required mathematical vocabulary that we use extensively in the following sections, and give a glimpse at how those theories can help deepen our understanding of the structure of the cosmic web. The reader interested in pursuing this investigation further should refer to the aforementioned references for a more detailed and involved introduction. In particular, we strongly recommend the reading of Gyulassy (2008) and especially Zomorodian (2009) for a very didactic presentation of these concepts. Indeed, the particular method and implementation presented in this paper are inspired by the work presented in those two references.

We then proceed by showing in section 5 how it is possible, relying on the previously mentioned theories, to design an algorithm that rigorously computes the *discrete* Morse complex of a discrete density field, obtained using DTFE technique (Schaap & van de Weygaert 2000) from the delaunay tessellation of a given discretely sampled data set, such as the distribution of galaxies in the universe. Within our approach, the Morse complex is directly computed from the delaunay tessellation which means it is scale adaptive and parameter free. The problem of dealing with Poisson noise and measurement errors is addressed in section 6, where we make use of persistence theory to remove spurious topological features from the Morse complex. Practically, the filamentary network (and associated voids, walls, ...) computed from the initial distribution is simplified by canceling pairs of critical points according to a persistence criterion, that can be restated in terms of significance relative to shot noise.



**Figure 1.** The dark matter density distribution in a  $50 h^{-1}$  Mpc large cosmological simulation (top left frame), with its ascending 3-manifolds (*i.e.* the voids, top right frame), ascending 2-manifolds (*i.e.* the walls, bottom left frame) and ascending 1-manifolds (*i.e.* the filaments, bottom right frame). The manifolds were computed using the method introduced in Sousbie et al. (2009).

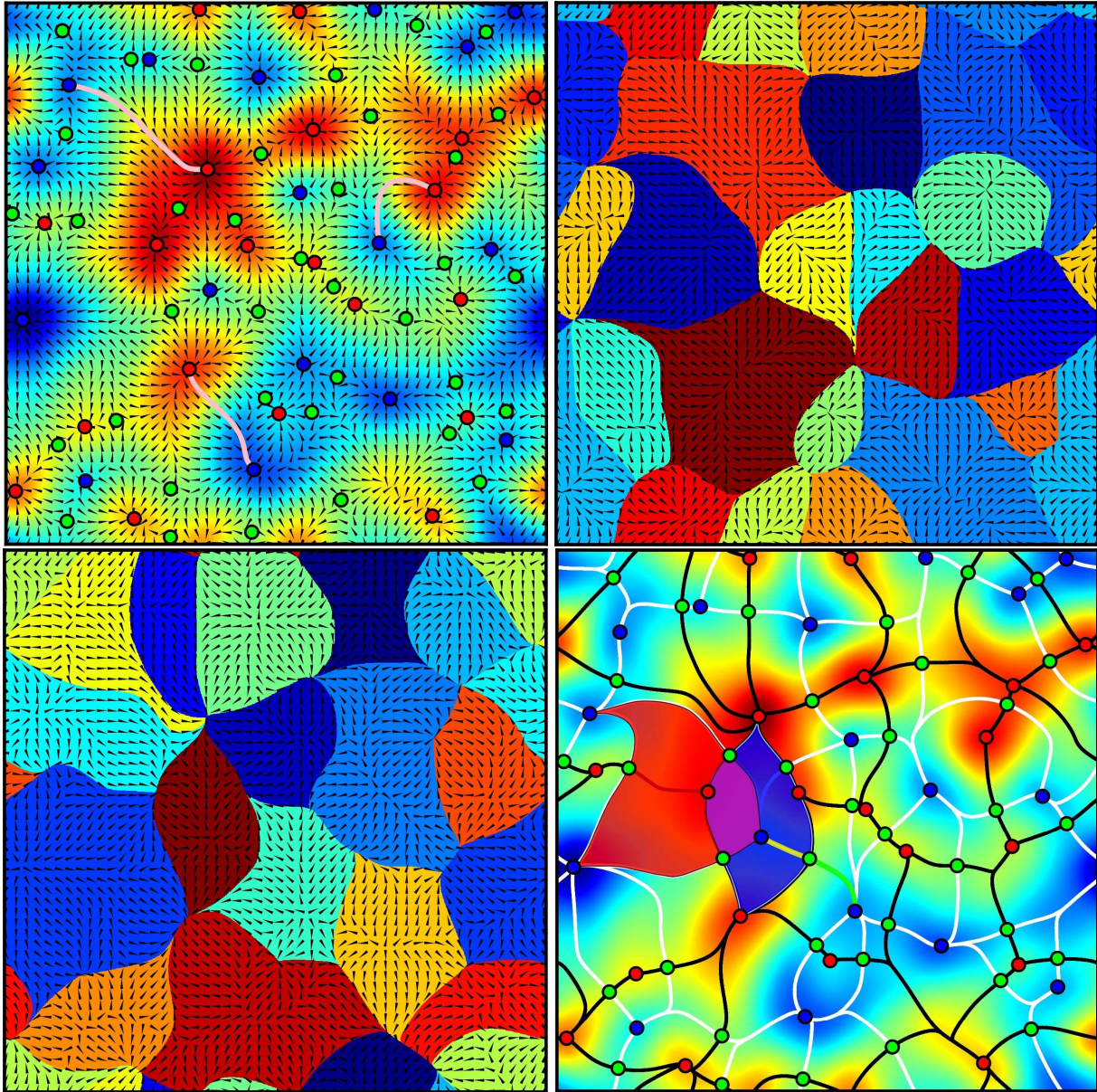
Finally, in section 7, we address technical questions such as dealing with boundary conditions, smoothing the identified voids, walls and filaments and important implementation problems before concluding in section 8.

Importantly, let us emphasize that within this framework, the mathematical theories that we use are fundamentally discrete and readily apply to the measured raw data; the unique supplementary but critical step consists in defining heuristically a consistent labeling of the segments, triangle and tetrahedron of the delaunay tessellation with regards to the DTFE densities computed at the sampling points (see section 5.1). This warrants that all the well known and extensively studied mathematical properties of the Morse complex are ensured *by construction* at the mesh level, and that the corresponding cosmological structures therefore correspond to well defined mathematical objects with known

mathematical properties. It also provides a consistent way of reconnecting the corresponding network after the removal of insignificant (non-persistent) pairs of critical points.

Note that a reference is given on the last two page, in which most mathematical terminology introduced in sections 2, 3 and 4 is defined in relatively simple terms. As we only aim here to introduce the necessary mathematical notions and giving a detailed description of the computation pipeline, extensively illustrating each step, the application to actual data sets is presented in a less technical companion paper, Sousbie, Pichon, Kawahara (2010). In that paper, we show the potential of this approach by applying it to typical cosmological data-sets: a large scale dark matter cosmological N-body simulation and the 7<sup>th</sup> data release (DR7) of the SDSS galaxy catalogue (Abazajian et al. 2009).





**Figure 2.** A 2D density field with its gradient (top left), its descending 2-manifolds (top right), its ascending 2-manifolds (bottom left), and its Morse-Smale complex (bottom right, see the black and white network). The maxima/saddle points/minima are represented as red/green/blue circled disks respectively and three integral lines are drawn in pink on the top left frame. On the central left part of the bottom right frame, an arc (*i.e.* a 1-cell) is represented in yellow (intersection of a green ascending 1-manifold and a blue descending 2-manifold) and a quad (*i.e.* a 2-cell) in purple (intersection of a red descending 2-manifold and a blue ascending 2-manifold).

## 2 MORSE THEORY FOR SMOOTH MANIFOLDS

Mathematically speaking, Morse theory is concerned with smooth scalar functions (say height of a mountain, or the temperature in a room) defined over generic manifolds. In the present case we are mainly interested in density fields: real valued functions defined over  $d$  dimensional Euclidian spaces<sup>1</sup>  $\mathbb{R}^d$ . We will therefore restrict the present discussion

to such geometries for the sake of simplicity. Morse theory provides a way to capture the intricate relation between the geometrical and topological properties of a function. What one means by geometrical property is basically any property unaffected by rigid motions such as translations or rotations. If  $h$  is the altitude function of a mountain landscape for instance, the altitude of the highest peak or its total surface are geometrical properties. Topology on the other hand captures how points are connected to each other with notions such as that of neighborhood. Topological properties are invariant under smooth continuous transformations. Sometimes topology is coined to be rubber geometry. Sticking to the landscape analogy and defining

<sup>1</sup> This is actually not generally true. Numerical simulations for instance often use periodic boundary conditions, which amounts to defining density on a torus  $\mathbb{T}^d \subset \mathbb{R}^d$ .

a mountain as the set of points that can be reached from its summit by going down the slope (*i.e.* following the gradient of  $h$ ), then the mountain itself is in some sense a topological property of the altitude function. Indeed, in winter, when covered with snow, or during summer, after the snow melted, the altitude map slightly changes, but the underlying mountain can still be easily identified as the same mountain. For the same reasons, a crest linking two mountains or a valley for instance are also topological properties of the landscape. When it comes to characterizing a function such as the matter density  $\rho$  on large scales in the universe, both topological and geometrical properties are interesting. While topological properties such as the number of galaxy clusters or dark matter haloes in a given volume are robust with respect to changes in the precise measured value of  $\rho$ , geometrical properties such as the density profile and precise location of a halo or a filament are more specific and characterize better the properties of  $\rho$ .

The relation between geometry and topology is intricate, and while modifying topology certainly requires a modification of geometry, the reverse is not generally true. For instance, the shape of a mountain may only slightly change with season, but more drastic events such as the explosion of a volcano (*i.e.* a drastic change in geometry) could actually erase it. Morse theory captures this relation for a generic function  $f$  by relying on the gradient  $\nabla_{\mathbf{x}}f(\mathbf{x}) = df/d\mathbf{x}(\mathbf{x})$  and its flow. The gradient defines a preferential direction at every point (the direction of steepest ascent) except where it vanishes (*i.e.* where  $\nabla_{\mathbf{x}}f = 0$ ). Those particular points are called critical points and can be classified according to the sign of the Hessian matrix, the  $d \times d$  matrix of the second derivatives  $\mathcal{H}_f(\mathbf{x}) = d^2f/dx_i dx_j(\mathbf{x})$ :

**Definition 2.1. (critical point of order  $k$ )** Let  $f$  be a function defined over  $\mathbb{R}^d$  and  $P$  a point with coordinate  $\mathbf{p} \in \mathbb{R}^d$ . Then  $P$  is a critical point of  $f$  if  $\nabla_{\mathbf{x}}f(\mathbf{p}) = 0$ . It is said to be of order  $k$  if the Hessian matrix  $\mathcal{H}_f(\mathbf{p})$  has exactly  $k$  negative eigenvalues.

Intuitively, in 2D, the top of a mountain is a maximum (order 2), a pass is a saddle-points (order 1) and the bottom of a valley a minimum (order 0). The top left frame of figure 2 shows the gradient and critical points of a function defined over  $\mathbb{R}^2$ . On this picture, the blue, green and red circles stand for the critical points of order 0 (minima), 1 (saddle points) and 2 (maxima) respectively. Note that according to definition 2.1, the order of a critical point is defined by the sign of the eigenvalues of the Hessian, which must therefore be non null. This condition is essential to Morse theory: a function  $f$  which obeys Morse theory must necessarily satisfy this constraint. Conversely, such functions are called Morse functions:

**Definition 2.2. (Morse function)** A Morse function is a smooth function whose critical points are non-degenerate. This means that for any  $P$  such that  $\nabla_{\mathbf{x}}f(\mathbf{p}) = 0$ ,  $\det \mathcal{H}_f(\mathbf{p}) \neq 0$ .

We will assume from now on that  $f$  is a Morse func-

tion.<sup>2</sup> At the location of any non critical point, the gradient indicates a preferred direction, and one can therefore define specific lines, the integral lines, by following the gradient flow:

**Definition 2.3. (Integral line or field line)** An integral line (also called field line) is a curve  $\mathbf{L}(t) \in \mathbb{R}^d$  such that

$$\frac{d\mathbf{L}(t)}{dt} = \nabla_{\mathbf{x}}f. \quad (1)$$

Its origin and destination are defined as  $\lim_{t \rightarrow -\infty} \mathbf{L}(t)$  and  $\lim_{t \rightarrow +\infty} \mathbf{L}(t)$  respectively.

The pink curves on top left frame of figure 2 show examples of integral lines: the lower order critical point at their extremity is their origin and the higher one their destination. The integral lines of a Morse function actually always have critical points as origin and destination. Let us consider the case of an integral line passing through a base point  $P$ . One can show that such integral line obeys certain properties:

**Property 2.3.1. (Integral lines of a Morse function)**

The integral lines of a Morse function  $f$  defined over  $\mathbb{R}^d$  and passing through a given point  $P$  is obtained by following the gradient and minus the gradient from  $P$ . It obeys certain properties:

- The origin and destination of an integral line is a critical point.
- Two integral lines passing through points  $P$  and  $P'$  respectively may only be identical or fully distinct : two integral lines cannot intersect (they can share their origin and/or destination though).
- The set of all the integral lines cover all of  $\mathbb{R}^d$  and each point  $P$  of space belong to exactly one integral line. It may be the origin/destination of several integral lines if it is a critical point though.
- An integral line with base point a critical point  $P$  is reduced to that point  $P$ .

The combination of the first and second properties is particularly interesting, as it allows classifying each points of space according to the origin *or* destination of its (unique) integral line. Such classification defines distinct regions of space called ascending and descending manifolds:

**Definition 2.4. (Ascending/Descending  $n$ -manifold)**

Let  $P$  be a critical point of order  $k$  of the Morse function  $f$  defined over  $\mathbb{R}^d$ . The ascending  $(d-k)$ -manifold defines a region of space with dimension  $(d-k)$ : the set of points reached by integral lines with origin  $P$ . The descending  $k$ -manifold defines a region of space with dimension  $k$ , the set of points reached by integral lines with destination  $P$ .

There exist exactly  $d$  different classes of ascending and descending manifolds, classified according to the order of the critical point at their origin or destination. Note that an ascending or descending  $d$ -manifold of a Morse function always spans a domain of dimension  $d$  (*i.e.* a 0-manifold is a (critical) point, a 1-manifold a line, a 2-manifold a surface,

<sup>2</sup> this is a strong requirement in practice, as shown in appendix A .



a 3-manifold a volume, ...). The central frames of figure 2 show the ascending and descending 2-manifolds of the 2D function on the upper frame. The notions of ascending and descending manifolds are actually at the core of Morse theory and the set of the descending or ascending manifolds is usually called the Morse complex<sup>3</sup>:

**Definition 2.5. (Morse complex)** the Morse complex of a Morse function  $f$  is the set of its ascending (or descending) manifolds.

The notion of Morse complex can actually be extended by following Smale and adding one more condition to a Morse function:

**Definition 2.6. (Morse-Smale function)** A Morse-Smale function is a Morse function whose ascending and descending manifolds intersect only transversely,

where the word “transverse” can be understood as the opposite of “tangent”, in the sense that there exist no point where two transverse manifolds are tangent. In other words, two ascending and descending manifolds should not be tangent and they should always penetrate into each other where they cross (*i.e.* they should “distinctly” intersect where they do). This additional condition ensures that the intersection of the ascending and descending manifolds is properly defined everywhere, so that the intersection of a  $p$ -ascending manifold and a  $q$ -ascending manifold may only have dimension  $n = \min(p, q)$  or be void. Such a non-null intersection is called a Morse-Smale  $n$ -cell:

**Definition 2.7. (Morse-Smale  $n$ -cell)** A Morse-Smale  $n$ -cell is the non void intersection of a  $p$ -ascending and a  $q$ -ascending manifold of a Morse-Smale function such that  $n = \min(p, q)$ . A 1-cell is generally called arc, a 2-cell is a quad and a 3-cell a crystal.

A  $n$ -cell is a refinement of the concept of an ascending/descending manifold. Whereas the descending and ascending manifolds are defined by the sets of integral lines having common origin *or* common destination respectively, a  $n$ -cell is defined by the sets of integral lines with common origin *and* destination. The bottom right frame of figure 2 displays examples of  $n$ -cells in 2D. The purple region for instance is the quad defined by the intersection of the red descending 2-manifold and the blue ascending 2-manifold: all integral lines within this region have the minimum on its boundary as origin and the maximum as destination (see also the upper right and lower left frames). Similarly, the yellow curve defines an arc at the intersection of the blue ascending 2-manifold and the green descending 1-manifold, as only one integral line has the minimum and saddle point at its extremities as origin and destination. The set of all  $n$ -cells defines the Morse-Smale complex:

**Definition 2.8. (Morse-Smale complex)** The Morse-Smale complex of a Morse-Smale function  $f$  is the set of all the  $n$ -cells of  $f$ .

On the same picture, the Morse-Smale complex is described by the critical points and the black and white curves. Basically, the critical points are its 0-cells, the set of black or white curves linking two critical points are its arcs (1-cells) and the regions bounded by a black and a white border are its quads (2-cells). In the 3D case we will consider in the next sections, the Morse Smale complex is also composed of 3-cells (the so called cristals). Note that the notion of  $n$ -cell is very interesting as it defines a natural partition of space induced by the flow of the gradient, literally dividing it into a so-called cell complex (a generalization of the concept of simplicial complex presented in section 3). We do not give further details here though as only the concept of arc is really needed for our purpose, the arcs really defining how critical points are connected to each other by integral lines. Actually, and although this is not formally correct, the reader may find it simpler to only consider the nodes (critical points) and arcs of the Morse-Smale complex complex, each arc connecting the critical points at their extremities, two critical points being potentially connected only if their order differ by 1 (*i.e.* a minimum and a 1-saddle, a 1-saddle and a 2-saddle, or a 2-saddle and a maximum). For instance, the arcs connecting maxima to saddle points are sub-sets of the ascending 1-manifolds and they enclose the information on how each filament (represented by its saddle point) connects exactly two maxima. Note that the geometry of an arc is determined by the integral lines whose origin and destination are the two critical points the arc connects. The Morse-Smale complex obey the following “combinatorial”<sup>4</sup> properties:

**Property 2.8.1. (Morse-Smale complex arcs)** the arcs (*i.e.* 1-cells) in the Morse-Smale complex connect critical points in such a way that:

- two arcs may only intersect at a critical point,
- an arc in the Morse-Smale complex links two critical points with index difference 1,
- there are exactly two descending arcs reaching a given critical point of order 1 (each departing from not necessarily distinct minima),
- there are exactly two ascending arcs departing from a given critical point of order  $d - 1$  (each reaching not necessarily distinct maxima).

Figure 1 illustrates how the theoretical concepts of Morse theory apply to cosmology. On this figure, the dark matter density distribution in a cosmological simulation is displayed on the top frame, together with its 3, 2 and 1 ascending manifolds on the second, third and fourth frame from top respectively. The ascending 3-manifolds associated to minima clearly trace the under dense regions usually denominated voids. The type 1 critical points trace the geometry of the walls through their ascending 2-manifolds and the filaments are traced by the ascending 1 manifolds, associated to critical points of type 1. As stated at the beginning of this section, Morse complex actually establishes the link between

<sup>3</sup> weather one chooses to use the ascending or the descending manifolds is only a matter of convention, as the descending  $n$ -manifolds of  $f$  are the ascending  $n$ -manifolds of  $-f$ .

<sup>4</sup> in this context, the term *combinatorial* is used to signify the discrete properties of the network formed by the Morse-complex: its number of nodes, their types, the number of branches and cycles, see below.

the geometrical (where are the critical points? what path does each arc follows?) and topological (how are critical-points connected? how many of each type are there?) properties of the cosmic web. Supposing the large scale matter density distribution  $\rho$  were a Morse function, each critical point of  $\rho$  could in fact be associated to a topological feature of the cosmic web whose geometry would then described by an ascending or descending manifold, the arcs defining a hierarchical neighborhood relation between them (the so called combinatorial property). The purpose of this paper is to construct, from the particles, a discrete Morse function which closely resemble<sup>5</sup> the sampled density (which in fact match it at the vertex of the tessellation) and which will therefore warrant all the corresponding discrete topological features.

### 3 DISCRETE MORSE THEORY

Even though the idea of applying Morse theory directly to the analysis of the cosmic web is quite appealing a priori, the task is actually not straightforward in practice. Indeed, Morse theory is defined for a Morse function, which is basically a smooth and at least twice differentiable real valued function satisfying the Morse criterion (definition 2.2). Whether it is because they result from fundamentally discrete processes, as in the case of galaxy distribution, or obtained through sampling, as for numerical simulation or observational data, typical astrophysical data sets typically do not comply to those criteria in general. In contrast, Discrete Morse theory, first introduced by Forman (1998b, 2002), is a very powerful theory which manages to capture the essence of the smooth Morse theory while still being readily applicable to discrete or sampled data commonly available to scientists. It is basically a combinatorial adaptation of Morse theory that applies to intrinsically discrete functions defined over simplicial complexes<sup>6</sup>.

Let us start by defining the basic building block of such spaces, the simplex. A  $k$ -simplex is the simplest possible geometrical figure of dimension  $k$ , or simply speaking the  $k$  dimensional analog of a triangle: A 0-simplex for instance is a point, a 1-simplex a segment, a 2-simplex a triangle, a 3-simplex a tetrahedron, ... More formally:

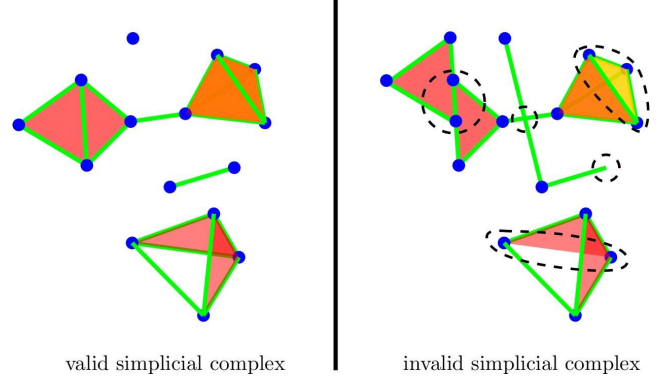
**Definition 3.1. ( $k$ -simplex)** A  $k$ -simplex  $\sigma_k$  is the convex hull of  $k + 1$  affinely independent points  $S = \{p_0, \dots, p_k\}$ . In other words, it is the set of points within the smallest possible solid with summits the  $k + 1$  points in  $S$ . It may be noted  $\sigma_k = \{p_0, \dots, p_k\}$ .

A simplex may have faces and cofaces:

**Definition 3.2. (face/coface of a  $k$ -simplex)** A face of a  $k$ -simplex  $\sigma_k$  with vertex  $S = \{p_0, \dots, p_k\}$  is any  $l$ -simplex  $\gamma_l$  with  $l \leq k$ , such that its vertex  $P = \{p_0, \dots, p_l\} \subset S$ . If

<sup>5</sup> Conversely, this construction would bias the reconstructed Morse-Smale complex if the underlying density was far from being a Morse function.

<sup>6</sup> actually, discrete Morse theory applies to the broader class of topological spaces called CW-complexes, which also include functions sampled over a regular cubic grid for instance.



**Figure 3.** Illustration of two sets of 3D simplexes,  $K$  and  $K'$ , forming a valid (left) and an invalid (right) simplicial complex. It is invalid because, from left to right and top to bottom, the intersection of the two 2-simplexes is not a simplex in  $K'$ , two 1-simplexes intersect, a 3-simplex (light yellow mostly hidden tetrahedron), a 1-simplex and a 2-simplex each lack one of their facets.

$\gamma_l$  is a face of  $\sigma_k$ , then  $\sigma_k$  is a coface of  $\gamma_l$ . In general, when  $k$  and  $l$  only differ by 1, a face is called a facet and a coface is called a cofacet.

Simply speaking, considering a tetrahedron in 3D (*i.e.* a 3-simplex) with 4 vertices as summits, its 2-faces are four triangles (*i.e.* its facets, any possible combination of three vertices), its 1-faces are 6 segments (*i.e.* any possible combination of two vertices) and its 0-faces are four points (*i.e.* any possible combination of one vertex). Reciprocally, the tetrahedron is a coface of any of those triangle, segments or points, and in particular it is a cofacet of any of the triangles. In general, a  $k$ -simplex has  $C_{l+1}^{k+1}$  faces of dimension  $l$ . Finally, a simplicial complex is a set of  $k$ -simplexes that comply to specific criteria:

**Definition 3.3. (simplicial complex)** A simplicial complex  $K$  is a finite union of simplexes such that

- Any face of a simplex in  $K$  also belongs to  $K$ .
- The intersection of two simplexes in  $K$  is empty or a simplex of dimension lower or equal, to the highest dimensional simplex they share.

Figure 3 shows an example of a combination of simplexes that form a simplicial complex (left frame) and a different combination that do not (right frame). A common example of a simplicial complex in astrophysics is the delaunay tessellation (see *e.g.* Okabe 2000; Schaap & van de Weygaert 2000) of a set of discretely sampled points.

As stated previously, discrete Morse theory directly applies to functions defined over a simplicial complexes. Those particular functions are called discrete functions, and for discrete Morse theory to apply, they also need to comply certain criteria:

**Definition 3.4. (Discrete Morse function)** A discrete function  $f$  defined over a simplicial complex  $K$  associates a real value  $f(\sigma_k)$  to each simplex  $\sigma_k \in K$ . The discrete function  $f$  is a discrete Morse function if and only if, for each  $\sigma_k \in K$ ,

- (i) there exist *at most* one facet  $\alpha_{k-1}$  of  $\sigma_k$  such that

$$f(\sigma_k) \leq f(\alpha_{k-1}),$$

(ii) there exist *at most* one cofacet  $\beta_{k+1}$  of  $\sigma_k$  such that  $f(\sigma_k) \geq f(\beta_{k+1})$ .

In other words, the Hessian non-degeneracy condition of smooth Morse theory (definition 2.2), becomes a condition on the value of the functions in the discrete theory: locally, a simplex has a higher value than its facets and a lower value than its cofacets, and there can only be one exception at most in each case. The reason for such a condition is not obvious at first sight but it is actually essential to the existence of a discrete gradient, the counterpart of the gradient in the smooth theory. In fact, if condition (i) and (ii) of the definition 2.2 of a discrete Morse function are satisfied then, locally, the discrete gradient of  $f$  (see below) can only define at most one preferential direction, as does the gradient of the corresponding smooth theory. Following this line of thought, the analog of a critical point of order  $k$  (see definition 2.1), a critical  $k$ -simplex of  $f$ , is a simplex for which  $f$  does *not* have any preferential relationship with one of its direct neighborhood (*i.e.* its facets and cofacets):

**Definition 3.5. (Critical  $k$ -simplex)** A  $k$ -simplex  $\sigma_k$  is critical for the discrete Morse function  $f$  if

(i) there exist *no* facet  $\alpha_{k-1}$  of  $\sigma_k$  such that  $f(\sigma_k) \leq f(\alpha_{k-1})$ ,

(ii) there exist *no* cofacet  $\beta_{k+1}$  of  $\sigma_k$  such that  $f(\sigma_k) \geq f(\beta_{k+1})$ .

It is important here to realize that the equivalent in discrete Morse theory of a critical point of order  $k$  is a critical  $k$ -simplex: in 2D, a minimum is a critical vertex (0-simplex), a saddle-point is a critical segment (1-simplex) and a maximum is a critical triangle (2-simplex).

Moreover, one can show that if definition 3.4 is satisfied, then at least one of the two conditions of definition 3.5 is verified, which leaves only two possible configurations for a simplex  $\sigma_k$ : exactly one of its cofacets and all its facets have a lower value or exactly one of its faces and all its cofacets have a higher value. In both cases, a preferential relation is established between  $\sigma_k$  and one of its facets or cofacets, which also defines a preferential direction, and leads to the following definition:

**Definition 3.6. (Discrete gradient vector field)** A discrete gradient vector field can be defined for a discrete Morse function  $f$  over  $K$  by coupling simplexes in gradient arrows (also called gradient pairs):

- if a simplex  $\sigma_k$  has exactly one lower valued cofacet  $\alpha_{k+1}$ , then  $[\sigma_k, \alpha_{k+1}]$  form a gradient arrow,
- if a simplex  $\sigma_k$  has exactly one higher valued facet  $\beta_{k-1}$ , then  $[\sigma_k, \beta_{k-1}]$  form a gradient arrow,
- if a simplex  $\sigma_k$  satisfies definition 3.5, it is critical, and does not belong to a gradient arrow.

Note that other configurations are impossible precisely because  $f$  is a discrete Morse function. Also, within a gradient arrow, the lowest valued simplex is the tail and the highest valued one the head (*i.e.* the discrete gradient actually points in the opposite direction of its smooth counterpart).

Figure 4 shows a discrete Morse function defined over a 2D simplicial complex (upper left frame), and its corresponding discrete gradient vector field and critical simplexes (upper right frame). One can note the similarity in the relation between the discrete gradient flow and the critical simplexes and that between the gradient and critical points on top frame of figure 2. Finally, one last important definition is that of the discrete integral line. In the terminology of Forman (1998a), it is called a V-path:

**Definition 3.7. (V-path)** A V-path is a strictly decreasing alternating sequence of  $k$ -simplexes  $\alpha_k^i$  and  $(k+1)$ -simplexes  $\beta_{k+1}^j$

$$\alpha_k^0, \beta_{k+1}^0, \alpha_k^1, \beta_{k+1}^1, \dots, \alpha_k^n, \beta_{k+1}^n,$$

where each pair  $\{\alpha_k^i, \beta_{k+1}^i\}$  forms a gradient pair and  $\alpha_k^{i+1}$  is a facet of  $\beta_{k+1}^i$ .

Tracing a V-path basically consists in intuitively following the direction of the gradient pairs, as one can see on the lower left frame of figure 4 where two V-pathes are highlighted in purple.

Using the previously introduced concepts, it becomes relatively straightforward to define a discrete Morse-Smale complex, and contrary to the smooth case, no manifold transversality condition (definition 2.6) needs to be enforced, as this is naturally achieved by the tessellation itself. In fact, following definition 2.4:

**Definition 3.8. (Discrete A./D.  $n$ -manifold)** Let  $\sigma_k$  be a critical simplex of order  $k$  of the discrete Morse function  $f$  defined over a simplicial complex  $K$ . The discrete ascending  $(d-k)$ -manifold is the set of  $k$ -simplexes that belong to at least one V-path with origin  $\sigma_k$ . The discrete descending  $k$ -manifold is the set of  $k$ -simplexes reached by field lines with destination  $\sigma_k$ .

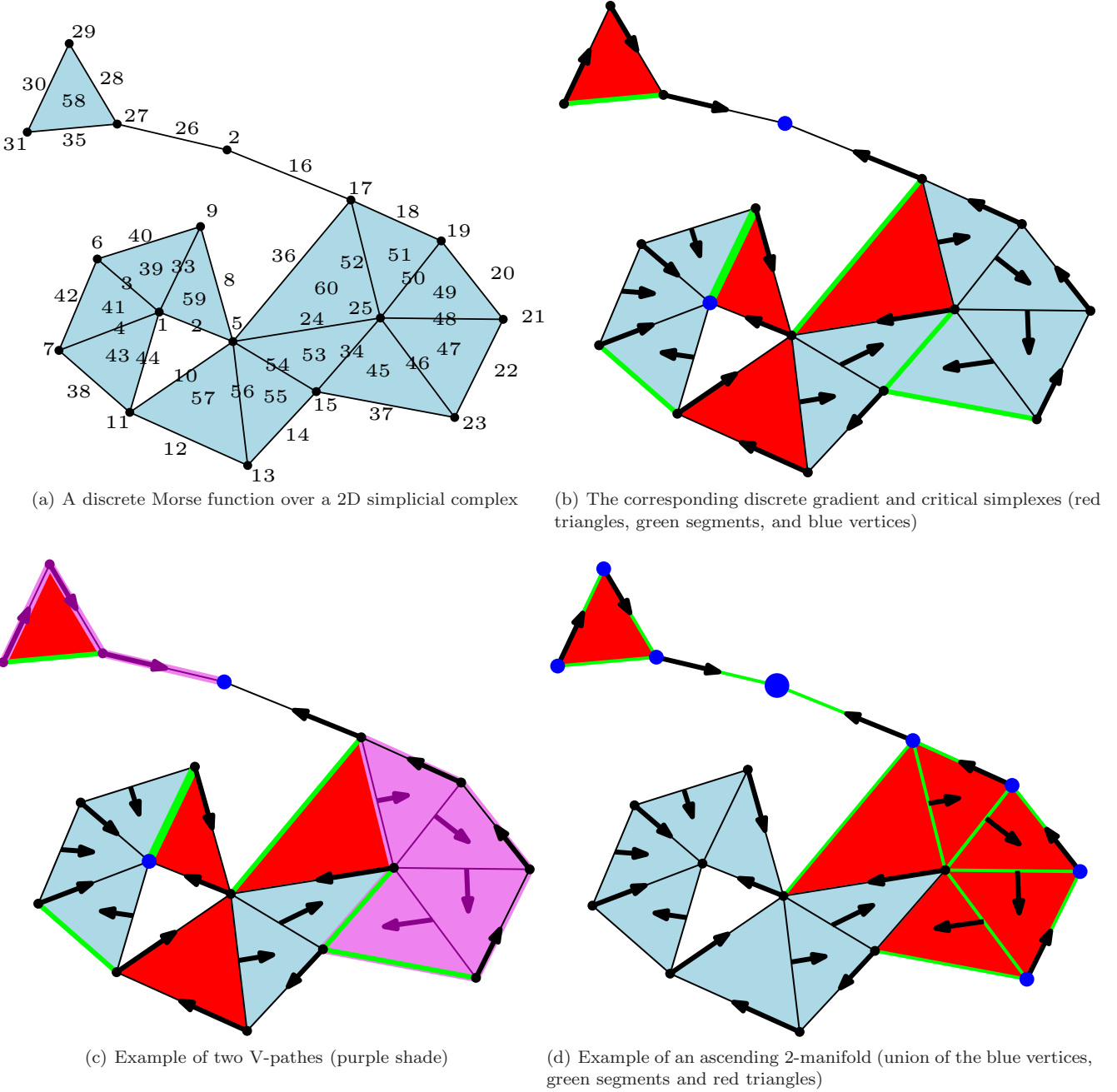
Note that according to that definition, a discrete  $k$ -manifold only contains  $k$ -simplexes (those in the V-pathes of  $\sigma_k$ ). This makes it difficult to define discrete  $n$ -cells (see definition 2.7) by intersecting manifolds, as they are made of simplexes with different dimensions. Following Gyulassy (2008), this definition is therefore extended to:

**Definition 3.9. (Extended Discrete A./D.  $n$ -manifold)** An extended discrete ascending (resp. descending)  $n$ -manifolds is a discrete ascending (resp. descending)  $n$ -manifold, together with its cofaces (resp. faces) and their extended discrete ascending (resp. descending)  $n$ -manifolds.

This literally fills lower dimensional “holes” in the manifold, making the intersection of two extended manifolds a very simple operation. On the lower right frame of figure 4 for instance, the discrete ascending 2-manifold is represented by the blue dots only. Their cofaces, the green segments, are included in the extended manifold, as well as their extended ascending manifolds (red triangles). The definition of the discrete Morse complex is therefore similar to the one in the smooth case:

**Definition 3.10. (Discrete morse complex)** the discrete Morse complex of a Morse function  $f$  is the set of its extended ascending (or descending) manifolds.





**Figure 4.** Illustration of the notions introduced by discrete Morse-theory. On the upper left panel (figure 4(a)), the numbers associated to each  $k$ -simplex (*i.e.* vertexes, segments and triangles) of the underlying simplicial complex define a discrete Morse-function. Note that a discrete Morse function must comply to definition 3.4, which is relatively restrictive, and in the present case, the function has been designed to illustrate notions of discrete Morse theory on a relatively small complex. We show in section 5.1 how a discrete Morse function can be defined to mimic the properties of a smooth function (such as the density or an altitude field for instance). The corresponding discrete gradient (see definition 3.6) is represented by the arrows on the upper right frame (figure 4(b)), each arrow associating a  $k - 1$ -simplex (the tail) to a  $k$ -simplex (the head). On the same frame (see also figure 4(c)), the red, green and blue shaded simplices are the critical 2, 1 and 0-simplices of the discrete function respectively (*i.e.* the equivalent of the maxima, saddle-points and minima of smooth theory). On the lower left frame (figure 4(c)), the two purple shaded sets of simplices correspond to two V-paths of the discrete Morse function (the discrete analog of an integral line, see definition 3.7). Intuitively, a V-path is a set of simplices linked by discrete gradient arrows, similarly to the integral lines of the smooth theory. Finally, the extended ascending manifold (see definition 3.9) of the minimum with value 2 (the large blue disk) is shown on the lower right frame (figure 4(d)). Similarly to the smooth theory, the corresponding ascending 0-manifold (definition 3.8) is defined by the set of simplices that one can reach by following the gradient arrows from the minimum (*i.e.* all the blue vertexes and green segments that belong to a gradient pair - *i.e.* an arrow - ). For the sake of consistency, one needs to define discrete extended manifolds (definition 3.9), which also include recursively the cofaces of any simplex in the discrete manifold, as well as the ascending manifolds of those cofaces that are critical. The resulting discrete extended ascending 0-manifolds is the set of blue vertexes, green segments and red triangles on the bottom right frame.

Similarly, a discrete  $n$ -cell is the intersection of two extended ascending and descending discrete manifolds (definition 2.7), and the discrete Morse-Smale complex remains the set of the discrete  $n$ -cells (definition 2.8). As in the smooth case, the discrete Morse-Smale complex is really a combinatorial object as it describes a particular way of grouping critical simplexes in pairs, quads, crystals ..., associating to each of those combination the geometry spanned by intersections of ascending and descending manifolds. We conclude by noting that, neglecting the effect of boundary conditions, the arcs of the discrete Morse-Smale complex (*i.e.* the V-pathes linking critical simplexes) obey the same properties as those of the Morse-Smale complex (definition 2.8.1).

#### 4 TOPOLOGICAL PERSISTENCE

The concept of persistence was first formalized in Edelsbrunner et al. (2002) (see also Robins (2000)). It is basically a method to quantify the importance of the topological features of a space, and was initially developed as a way to robustly measure topological properties when noise is present, and to enable topological simplification (*i.e.* the modification of a function or a space so that its less significant *topological* features are removed). The theory was originally described in the context of simplicial homology (for functions defined over a simplicial complex, see appendix B) and was very nicely exposed in Edelsbrunner et al. (2002). Let us stress here that the concept of persistence itself is largely independent of the fact that a function is smooth or not, as it only quantifies the robustness of its topological properties given one can measure them, whatever the nature of the function itself. Let us illustrate here the idea behind the concept of persistence using simple examples.

For smooth functions, persistence theory is based on the evolving properties of the so-called sub-level sets of a function  $\rho$ , as they change with the value of the level  $\rho_0$ . A sub-level set is the set of points where  $\rho(\mathbf{x} = (x_1, \dots, x_n))$  is higher than or equal to a certain value  $\rho_0$ :

**Definition 4.1. (Level set and Sub-level set)** A level set (also called isocontour) of a function  $\rho(\mathbf{x})$  of  $n$  variables  $x_i$  at level  $\rho_0$  is defined as

$$(x_1, \dots, x_n) | \rho(x_1, \dots, x_n) = \rho_0.$$

A sub-level set (also called excursion set) is defined as

$$(x_1, \dots, x_n) | \rho(x_1, \dots, x_n) \geq \rho_0.$$

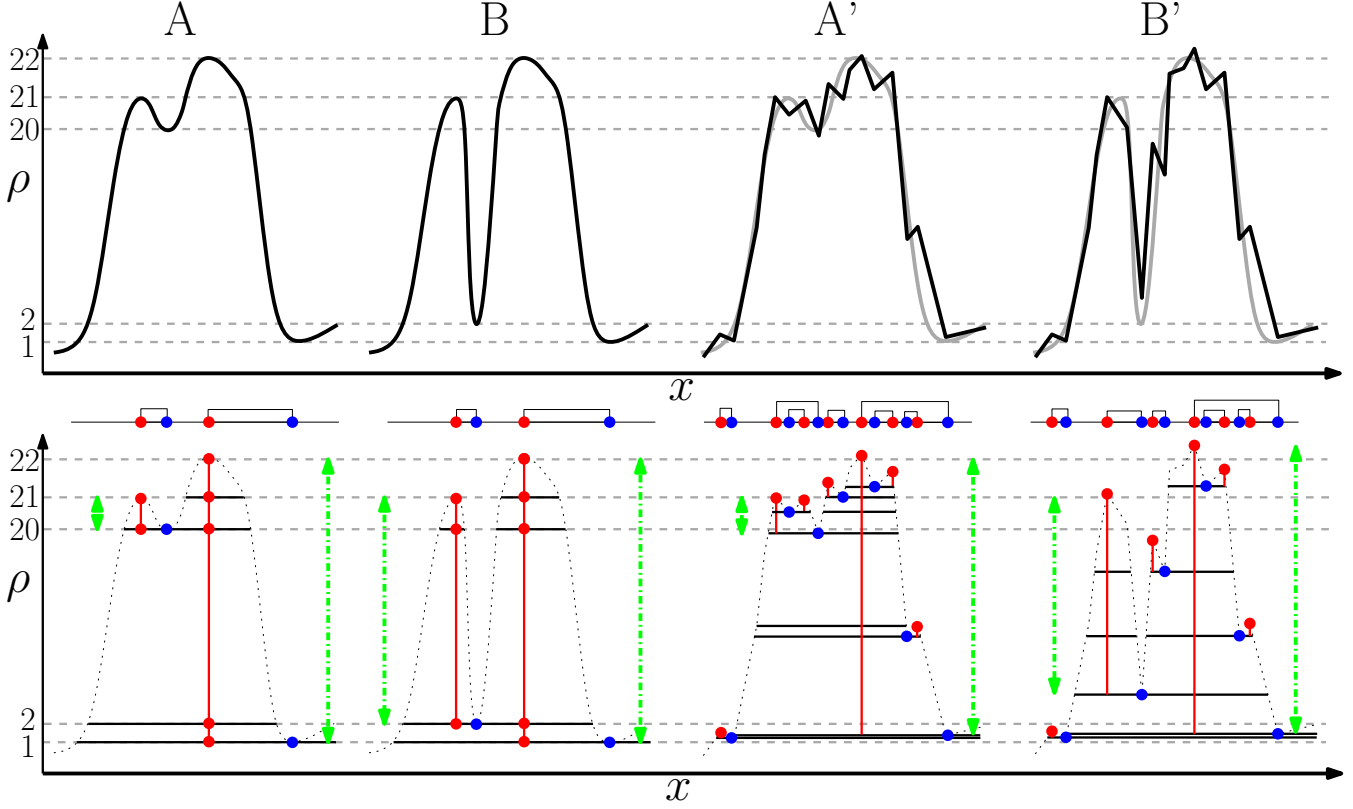
Using this definition, persistence can be interpreted as a measure of the “life-time” of topological features, the so called  $k$ -cycles, in the sub-level sets. When the value of  $\rho_0$  skims through the image of  $\rho$  (*i.e.* the set of values  $\rho(\mathbf{x})$  may take), the corresponding sub-level set grows, and the way it is connected evolves. In 3D, isolated islands (also called components or 0-cycles) first appear around the maxima. Those islands later merge into each other at saddle points of type 1 to finally form rings bordering holes (the 1-cycles). For lower values of  $\rho_0$ , those holes get filled at saddle points of type 2, destroying the corresponding 1-cycles, to later form spherical shells around minima (the 2-cycles), when a sufficient number of holes have been filled

and those spherical shell also end up being filled at minima, therefore destroying the corresponding 2-cycles. Persistence then relates the importance of a given  $k$ -cycle to the length of the interval of values  $\rho_0$  can take and for which a given  $k$ -cycle exists within the growing sub-level sets.

Figure 5 illustrates how persistence works in 1D. On this figure, the upper part displays four different functions, where the two on the right (labeled  $A'$  and  $B'$ ) were obtained by discretely sampling the two on the left (labeled  $A$  and  $B$ ), adding random noise, and linearly interpolating between the sample points. The lower part of the figure shows the different sub-level sets of these functions for values corresponding to their critical points. On the bottom left frame for instance, the sub-level sets of  $\rho(x)$  are empty for levels  $\rho_0 > 22$ . At level  $\rho_0 = 22$  though, a new component (*i.e.* a 0-cycle) appears, which corresponds to the highest maximum of the function. This component grows for levels  $22 > \rho_0 > 21$  and a new independent components appears at the level of the second highest maximum,  $\rho_0 = 21$ . Those two components remain independent while  $\rho_0 > 20$  but merge when reaching  $\rho_0 = 20$ , the value of the first minimum. Basically, the minimum *destroyed* a component that was *created* by a maximum. By convention, we say that it destroys the most recently created one (*i.e.* the maximum with lowest density), and that the minimum and left maximum therefore form a persistence pair (as illustrated on the central sketch) with persistence  $21 - 20 = 1$ . The four sketches on the bottom part illustrate this pairing process for the four different function. One should note that a given critical point may not always be paired in the process, and that because the 1D case is very simple, a given type of critical point always create or always destroy, but this is not the case in general, for critical points that are not extrema.

A very common task when studying galaxy distributions or cosmological N-body simulations involves identifying galaxy clusters or dark matter haloes. This is often achieved using relatively simple but robust methods, such as the friend-of-friend algorithm (Huchra & Geller 1982), that basically involve carefully selecting a global level  $\rho_c$  and considering each independent component in the sub-level set  $\rho_c$  of the density field  $\rho(\mathbf{x})$  as an independent cosmological structure. Applied to the functions  $A$  of figure 5 for example, such a method may detect one or two different peaks with  $\rho_c = 19$  or  $\rho_c = 20.5$  respectively, but it will not yield any information on whether those peaks are comparable or if one of them is more meaningful than the other. Persistence on the other hand can make such distinction, because it is built using information present in *all* the sub-level sets: while function B contains two comparably persistent peaks, function A really only contains one (the peaks persistence is symbolized by the length of the green arrows on the figure). The remarkable fact is that this stays true even if the sampling is poor and noise is present, as illustrated by function  $A'$  and  $B'$ . Because of noise, many spurious peaks exist in this two functions, which may potentially lead to numerous fake identifications, but even in that case, whereas a density selection method would clearly fail to count the peaks correctly, persistence easily identifies the presence of only one persistent peaks in  $A'$ ,





**Figure 5.** Illustration of the concept of persistence over a 1D functions. The upper panel shows two functions (left) and their discretely sampled counterparts, with noise added (right). The lower panel displays the evolution of sublevel-sets of these functions at the level of different critical points as one spans densities from high to low. The green dash-dotted vertical arrows emphasize the lifetime of components in the sublevel-sets, the persistence pairs are displayed in the central part over the function's Morse-Smale complex.

and two in  $B'$ , as in the case of functions  $A$  and  $B$ .

Although a very simple 1D case was illustrated here, the general idea remains the same in higher dimensional spaces. In general, one studies how components of sub-level sets are created or destroyed, but in higher dimensions, one also has to keep track of more complex structures than independent components, such as the formation of 2D holes or 3D shells in the structure (*i.e.* the 1-cycles and 2-cycles).

As was mentioned earlier, persistence can equally be computed directly for discrete functions defined over simplicial complexes, given that one can define a concept similar to that of growing sub-level sets in such context. Filtration is such a concept. A discrete function  $v$  associates a value to each simplex in a complex, and one can for instance define a filtration  $F$  of a simplicial complex  $K$  as the sets of sub simplicial complexes  $K_i$  such that only the simplexes  $\sigma$  with value  $v(\sigma) < v_i$  belong to each  $K_i$ . More generally,

**Definition 4.2. (Filtration)** A filtration of a finite simplicial complex  $K$  is a sequence of  $N + 1$  sub-complexes  $K^i$  of  $K$  such that:

- (1)  $\emptyset = K^0 \subseteq K^1 \subseteq \dots \subseteq K^{N-1} \subseteq K^N = K$ ,
- (2)  $K^{i+1} = K^i \cup \delta^i$ ,

where  $\delta^i$  is a subset of the simplexes in  $K$ , and  $A \subseteq B$  means that  $A$  is included in or equal to  $B$ . One can in particular define a special filtration induced by a function  $v$  that asso-

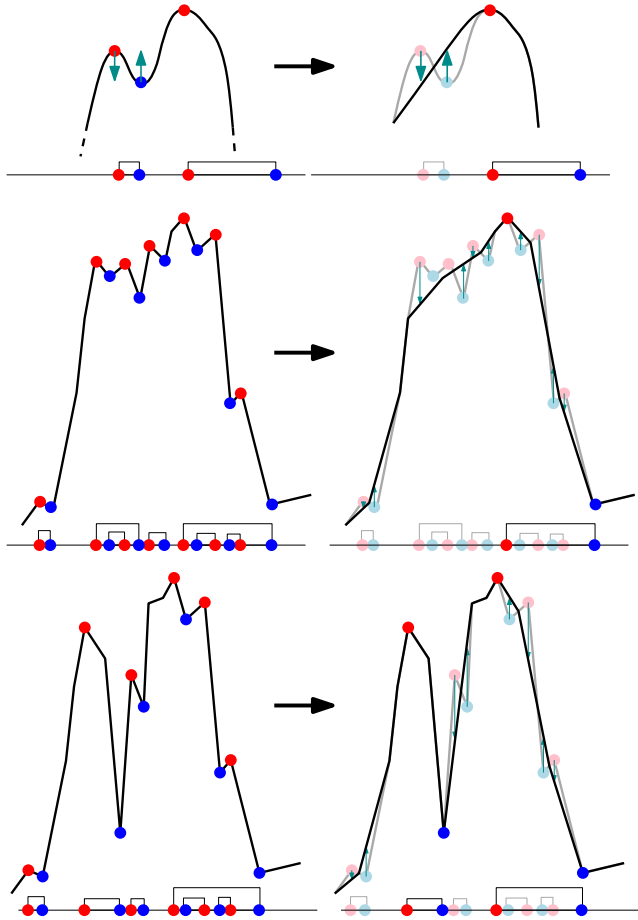
ciates a value to each  $\sigma \in K$ , such that each  $K_i$  is the set of simplexes in  $K$  with value  $v(\sigma)$  less or equal to a given threshold  $v_i$ .

In that case, each  $K_i$  is the discrete equivalent of the growing sub-level sets of definition 4.1 (see also figure C1 for an example of a filtration) for the discrete function  $v$  that defines the order of entrance of simplexes within the filtration. As the filtration grows with increasing value of  $v_i$ , new components, loops, shells, ... appear. As for the smooth function counterpart, those topological features are generally called  $k$ -cycles, and we define them formally in the context of discrete theory (this definition would conceptually be very close in the context of smooth functions though):

**Definition 4.3. ( $k$ -cycle)** a  $k$ -cycle in a simplicial complex  $K$  is a  $k$  dimensional topological feature with  $0 \leq k < D$ , where  $D$  is the number of dimensions. When  $D = 3$  for instance, a 0-cycle is an independent component (*i.e.* a set of simplexes non-linked to the rest of the complex), a 1-cycle is a loop (a set of simplexes that form a ring with a hole in the middle) and a 2-cycle is a shell (a set of simplexes bounding a 3D empty region).

As for a smooth function, one can therefore track the creation and destruction of  $k$ -cycles in  $F$  as simplexes enter the filtration, pairing critical simplexes into persistence pairs:

**Definition 4.4. (Persistence)** persistence measures the “life-time” of topological features (*i.e.*  $k$ -cycles) in a filtra-



**Figure 6.** Illustration of the topological simplification process applied to functions  $A$ ,  $A'$  and  $B'$  defined on figure 5 (see top, central and bottom panel). The diagram under each function represents its Morse-Smale complex and persistence pairs.

tion of a finite simplicial complex  $K$  induced by a discrete function  $v$  or equivalently in the growing sub-level sets of a smooth function  $\rho$ . The arrival of each critical simplex in the discrete case or critical points in the smooth case corresponds to the creation or destruction of a topological feature ( $k$ -cycle). Persistence pairs critical simplices  $\sigma_a - \sigma_b$  (or critical points  $P_a - P_b$ ) that create and destroy a given feature, their corresponding persistence being defined by the difference of their “arrival time”,  $v(\sigma_a) - v(\sigma_b)$  (or  $\rho(P_a) - \rho(P_b)$ ). It can also sometimes be useful to define a persistence ratio as the ratio of those values.

The computation of persistence pairs in a 2D filtration is illustrated in appendix C and intuitively, persistence describes how much a function would need to change to remove a topological feature.

The main interest of being able to identify persistence pairs of critical points (or simplices) in a given function is that it yields an objective topological criterion to assess the significance of those critical points (or simplices). Actually, one can go even further and show that it is actually always possible to *locally* modify the function to cancel non persistent pairs out and therefore remove topological noise. The process is illustrated on figure 6. In

1D, a persistence pair is always formed of a minimum and a maximum. If those two critical points are direct neighbors, one can in fact increase the value around the minimum and decrease the value around the maximum until the value at the maximum becomes smaller than that at the minimum. When this happens, both points are not critical anymore and none of the other critical points are affected. On the top panel for instance, the process is applied to the less persistent bump of function  $A$ . Note that the details of how the function is modified are not important; what is the fact that it is possible to cancel a non persistent pair and remove it from the Morse-complex (see the diagrams below the functions). For instance, if one considers that structures whose persistence is lower or equal to the persistence of the smaller bump of function  $A$  are not significant (*i.e.* generated by noise with high probability), then one can deduce cancel the corresponding topological features so that function  $A$  becomes topologically equivalent to its simplified version (top right) with the corresponding Morse complex. Applying the same process to  $A'$ , the noisy version of  $A$ , one actually obtains a function with identical topology and Morse complex (central panel). This means that even in the presence of a relatively important noise, it is still possible using persistence to recover the topology and Morse complex of the underlying function (see also the bottom panel to check how the topology of function  $B$  on figure 5 can be recovered from its noisy counterpart,  $B'$ ). We detail in section 6.2 a generic algorithm that implements symbolic topological simplification in order to recover the structure of the Morse complex of matter distribution on large scale from a raw noisy version computed directly over a Delaunay tessellation.

## 5 DISCRETE MORSE COMPLEX

The basis of the necessary mathematical background being introduced in sections 2, 3 and 4, we now start detailing the particular algorithm and implementation used in DisPerSE. As previously mentioned, our purpose is to compute a *discrete* Morse complex and use its properties to identify and characterize the structure of the cosmic web. This approach has both advantages of being applicable to spaces with 3 or more dimensions and having a solid mathematical framework see also Gyulassy (*e.g.* chap. 6, 2008) (see also Gyulassy (chap. 6 2008)). To summarize, a simplicial complex is computed from a discrete distribution (galaxy catalogue, N-body simulation, ...) using Delaunay tessellation and a density  $\rho$  is set to each galaxy using DTFE (roughly speaking, the density at a vertex is proportional to the inverse volume of its dual Voronoi cell, see Schaap & van de Weygaert (2000)). A discrete Morse function is then defined by heuristically tagging a properly chosen value to each simplex in the complex (*i.e.* the segments, facets and tetrahedron of the tessellation). From this discrete function, we then compute the discrete gradient and deduce the corresponding discrete Morse-Smale complex (DMC hereafter, see section 3; Forman (2002)). The DMC is then used as the link between the topological and geometrical properties of the density field. Its critical points together with their ascending and descending manifolds are identified



to the peaks, filaments, walls and voids of the density field (see section 2). At this stage, the DMC is mainly defined by Poisson sampling noise and measurement uncertainties, and we filter it using persistence theory (see section 4 and appendix B and C). For that purpose, we consider the filtration of the tessellation according to the values of the discrete Morse function and use it to compute persistence pairs of critical points. The DMC is finally simplified by canceling the pairs that are likely to be generated by noise. This is achieved by computing the probability distribution function of the persistence ratio of all types of pairs in scale invariant Gaussian random fields and canceling the pairs with a persistence ratio whose probability is lower than a certain level.

### 5.1 Discrete gradient

As stated in section 3, a discrete gradient field is derived from a proper discrete Morse function, which must satisfy definition 3.4. Although those conditions are restrictive enough to make the deduction of a valid discrete Morse function difficult, they allow for a wide variety of such functions to exist; one has to keep in mind that the final discrete gradient field should be as similar as possible to its continuous counterpart  $\nabla\rho$ , the gradient of the density field  $\rho$ . We therefore note the discrete Morse function  $F_\rho$ . An optimal method to define a discrete gradient has yet to be discovered, but Lewiner (2002) propose a nice review on the topic and relatively advanced solutions. Unfortunately, these solutions involve the computation of relatively complex hyper-graphs and are not easily applicable to large data sets. Instead, we implement here a modified version of the one presented in Gyulassy (2008), which present the advantage of not depending on an arbitrary labeling of the simplexes.

Let  $F_\rho$  be the discrete Morse function computed from a smooth function  $\rho$  over a simplicial complex  $K$ , with  $\sigma_k$  a  $k$ -simplex that belongs to  $K$ ,  $\text{Facet}[\sigma_k] \in S$  the facets of  $\sigma_k$  and  $\text{Vertex}[\sigma_k] \in S$  the faces of  $\sigma_k$  with dimension 0 (*i.e.* its vertices). The value of the discrete Morse function at  $\sigma_k$  is defined as:

$$\begin{cases} k > 0 : & F_\rho(\sigma_k) = \max(F_\rho(\text{Facet}[\sigma_k])) \\ & \quad + \epsilon^k \cdot \sum F_\rho(\text{Vertex}[\sigma_k]), \\ k = 0 : & F_\rho(\sigma_0) = \rho(\sigma_0), \end{cases} \quad (2)$$

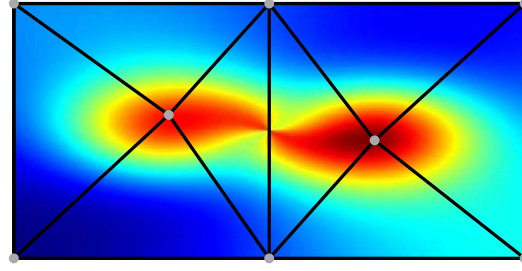
where  $\max()$  stands the maximal value of its arguments,  $d$  is the number of dimensions, and  $\epsilon$  is an infinitely small value. One can easily check that such a function does comply to the definition 3.4 of a discrete Morse-function. In fact, the value of a simplex is always slightly higher than the value of its highest facet, and thanks to the factor of  $\epsilon^k$ , two simplexes sharing the same highest facet have different values if two vertices in  $K$  cannot have the same density. In practice, this is always the case when computing densities using DTFE and in the following we will therefore assume that we are in such a situation. For that reason, equation 2 defines the value of  $F_\rho$  uniquely from a given smooth function  $\rho$ , and independently of any arbitrary

labeling of the simplexes. Note that to compute the DMC, one only needs to be able to compare simplexes, and it is therefore not necessary to give a particular value to  $\epsilon$ , as only a comparison operator needs to be implemented. This definition of  $F_\rho$  allows for a unique ordering over the simplexes of  $K$ .

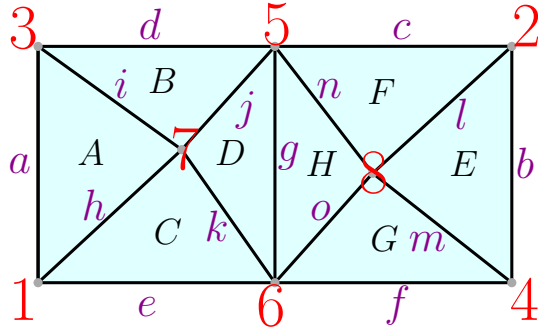
As explained in section 3, a discrete gradient can be defined over  $K$  by grouping pairs of simplexes whose dimension differ only by 1 (*i.e.* a vertex and a segment, a segments and a triangle or a triangle and a tetrahedron) and such that conditions 3.6 are satisfied. A group of two paired simplexes form a gradient pair, and the remaining unpaired simplexes are critical (the equivalent of the critical point for a smooth density field<sup>7</sup>). Looking at conditions 3.6, one can see that for two simplexes to form a gradient pair, the simplex of lower dimension should always have a value higher than the other. But because  $F_\rho$  has precisely been defined such that any simplex has a value higher than its facets, no pair may be formed, and all the simplexes in  $K$  are therefore initially critical. As a consequence, the Morse complex of  $F_\rho$  can be readily deduced: each  $k$ -simplex is a critical simplex of order  $k$ , and it is linked by an arc to each of its faces and cofaces, which are also critical. Many of those arcs actually link critical simplexes whose discrete Morse function  $F_\rho$  only differ by an infinitesimal amount  $\Delta F_\rho \propto \epsilon^p$  though, and we call such arcs  $\epsilon$ -persistent. Because along those arcs, the value of the function only changes infinitesimally, they can be canceled while only modifying the value of  $F_\rho$  by an infinitely small amount. In fact, doing so one can basically exchange the values of  $F_\rho$  given to each critical simplex at the extremity of the  $\epsilon$ -persistent arc and pair them within a gradient arrow. By repeating this process until no  $\epsilon$ -persistent arcs exist anymore, one can therefore deduce a correct discrete gradient.

In practice, we proceed by considering the sets of the  $k$ -simplexes of  $K$  one by one, in ascending order of their dimension, and within each set, we iterate over the simplexes  $\sigma_k$  in ascending order of their value  $F_\rho(\sigma_k)$ . For each of them, if it is not already in a gradient pair, we retrieve the lowest of its cofaces  $\alpha_{k+1} \in \langle \sigma_k \rangle$  that is not already in a gradient pair and which value is only infinitesimally higher than  $F_\rho(\sigma_k)$ . If it exists, we pair them and else,  $\sigma_k$  remains unpaired. Note once again that the value of  $F_\rho$  does not need to be explicitly modified in the actual implementation, as  $\alpha_{k+1}$  and  $\sigma_k$  may only differ infinitesimally if  $\sigma_k$  is the highest facet of  $\alpha_{k+1}$ . The algorithm ends when all the simplexes have been checked once. We show on figure 7 a practical example of how the algorithm runs on a simple smooth function and a 2D simplicial complex spanning over its domain of definition.

<sup>7</sup> Note that a critical point of type  $k$  from the smooth theory is equivalent to a critical  $k$ -simplex of the discrete theory. In 2D, minima are critical vertices, saddle-point critical segments and maxima are critical triangles.

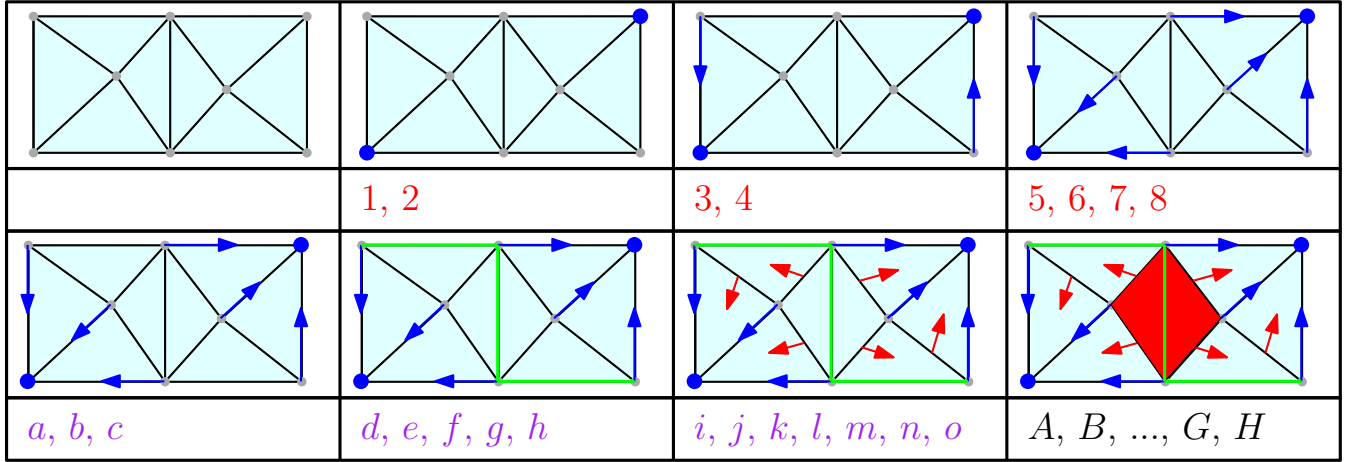


(a) Example of a smooth function and a simplicial tessellation of space



$$\begin{aligned}
 a &= 3 + 4\epsilon & e &= 6 + 7\epsilon & i &= 7 + 10\epsilon & m &= 8 + 12\epsilon \\
 b &= 4 + 6\epsilon & f &= 6 + 10\epsilon & j &= 7 + 12\epsilon & n &= 8 + 13\epsilon \\
 c &= 5 + 7\epsilon & g &= 6 + 11\epsilon & k &= 7 + 13\epsilon & o &= 8 + 14\epsilon \\
 d &= 5 + 8\epsilon & h &= 7 + 8\epsilon & l &= 8 + 10\epsilon \\
 A &= 7 + 10\epsilon + 11\epsilon^2 & E &= 8 + 12\epsilon + 14\epsilon^2 \\
 B &= 7 + 12\epsilon + 15\epsilon^2 & F &= 8 + 13\epsilon + 15\epsilon^2 \\
 C &= 7 + 13\epsilon + 14\epsilon^2 & G &= 8 + 14\epsilon + 18\epsilon^2 \\
 D &= 7 + 13\epsilon + 18\epsilon^2 & H &= 8 + 14\epsilon + 19\epsilon^2
 \end{aligned}$$

(b) The corresponding discrete Morse function defined over the simplicial complex (see equation 2)



(c) Computation of the discrete gradient

**Figure 7.** Illustration of the computation of a discrete gradient from a simple smooth function and a simplicial complex spanning its domain of definition, as shown on panel 7(a). The corresponding discrete Morse function is represented on panel 7(b). Each vertex is labeled with the value of the corresponding smooth function, and the lower case and upper letters correspond to the labels of the segments and triangles respectively, for which the corresponding value of  $F_\rho$  is shown on the right of the panel (see equation 2). Note that sorting segments or triangles labels according to alphabetical order also sorts them in increasing order of their value. Panel 7(c) illustrates the computation of the discrete gradient according to the algorithm described in section 5.1, which works by considering the vertexes, segments and triangles one after the other, in increasing order of their value (from left to right and top to bottom on the diagram). Starting with the first vertex,  $F_\rho^{-1}(1)$  (lower left vertex), its cofaces are the segments labeled *a*, *h* and *e* with value  $3 + 4\epsilon$ ,  $7 + 8\epsilon$  and  $6 + 7\epsilon$  respectively. As none of those value differ from 1 by a factor of  $\epsilon$  only, no pair can be formed, and the vertex remains critical (*i.e.* unpaired, represented by a blue disk on the diagram). The vertex with value 2 presents the same configuration, and is therefore also critical, but the third one to enter, labeled 3, has one available coface labeled *a* with value  $3 + 4\epsilon$  that is only infinitesimally higher, which means the vertex and segment form a gradient pair (blue arrow between 3 and *a* on the diagram). The case of vertex 4 is similar, and it is paired to segment *b*. The next vertex, labeled 5 is problematic because it presents two cofaces with infinitesimally higher value, *c* and *d*, but the conflict is easily solved by pairing with one with value closest from 5, segment *c*. We then proceed until no vertex is available anymore, and start considering segments (leftmost box of the second row on the diagram). Segments *a*, *b* and *c* are skipped because they are already paired to vertex 3, 4 and 5 respectively. Segment *d* is free though but does not have an infinitesimally higher coface (*i.e.* triangle *B*), it is therefore a critical segment (*i.e.* the equivalent of a saddle point, represent as in green). Segments *e* and *h* are paired while *f* and *g* are found to be critical. This leads to segments *i* whose cofaces are *A* and *B*, whose value differ from that of *i* by  $11\epsilon^2$  and  $15\epsilon^2$  respectively, *i* is therefore paired to the closest triangle in value, *A* (red arrow on the diagram). The remaining segments are processed the same way and one can then start reviewing the triangles. Only *D* and *H* are not paired, and as in 2D triangles have no cofaces, they are critical (colored red on the diagram). The final discrete gradient is shown on the bottom right box of the figure 7(c).



## 5.2 Discrete Morse complex computation

Once a proper discrete gradient has been defined over a simplicial complex, it becomes relatively straightforward to deduce its corresponding DMC. According to definition 2.4, the ascending (resp. descending) manifold of a critical point  $P_k$  of order  $k$  is the set of integral lines that end (resp. start) at  $P_k$ . The discrete analog of an integral line is a V-path (i.e. a sequence of simplexes linked by the discrete gradient, see definition 3.7) and one can therefore identify ascending (resp. descending) manifolds by following the V-paths that end (resp. start) at a critical simplex  $C_k$ . The core of the algorithm consists in a simple “breadth first search” where sequences of cofaces and gradient pairs are identified according to definition 3.7. Each manifold is stored in a separate set type data structure as one simplex may be reached by different V-paths within one manifold. Let  $\mathcal{A}(C_k)$  be the set that stores the ascending manifold of the critical  $k$ -simplex  $C_k$ . The recursive algorithm starts by considering the set of the *cofacets* of  $C_k$ , stored in an array  $A_{cur}$  that will basically contain, at the  $n^{\text{th}}$  step of the algorithm, the set of  $(k+1)$ -simplexes in the  $n^{\text{th}}$  gradient pair of any V-path starting at  $C_k$ . At each step, the content of  $A_{cur}$  is scanned and for each  $(k+1)$ -simplex, there exist four possibilities:

- (i) it is critical, in which case it is skipped as the V-path ends.
- (ii) it is not critical, and is paired to a  $k$ -simplex in a gradient pair. In that case, the  $k$ -simplex is added to  $\mathcal{A}(C_k)$  and stored in a temporary array  $A_{tmp}$ .
- (iii) it is not critical, but is paired by discrete gradient to a  $k$ -simplex already in  $\mathcal{A}(C_k)$ . In that case, it is skipped.
- (iv) it is not critical, and is paired to a  $k+2$ -simplex in a gradient pair. In that case, it is skipped.

Once all simplexes in  $A_{cur}$  have been treated, the content of  $A_{cur}$  is replaced by the *cofacets* of the  $k$ -simplexes in  $A_{tmp}$  and the process is iterated until  $A_{cur}$  is empty at which stage all the simplexes in  $\mathcal{A}(C_k)$  have been retrieved. The computation of the descending manifold  $D(C_k)$  is achieved the exact same way, except that cofacets are replaced by facets. A pseudo-code implementation is presented in Algorithm 1 (see the non tagged lines only). Note that in this implementation, only  $k$ -simplexes are stored to describe the manifold of a critical  $k$ -simplex, which reduces memory usage. It also implies that the algorithm does not compute the extended discrete manifolds of definition 3.9 but rather those of definition 3.8. This is indeed not a problem though as those manifolds can easily be extended at query time from the identified sets of  $k$ -simplexes. Practically, extending an ascending (resp. descending)  $k$ -manifolds consists in recursively adding the cofaces (resp. faces) of any simplex in the manifold, as well as the ascending (resp. descending)  $p$ -manifolds ( $p > k$ ) of any of its critical  $p$ -simplexes.

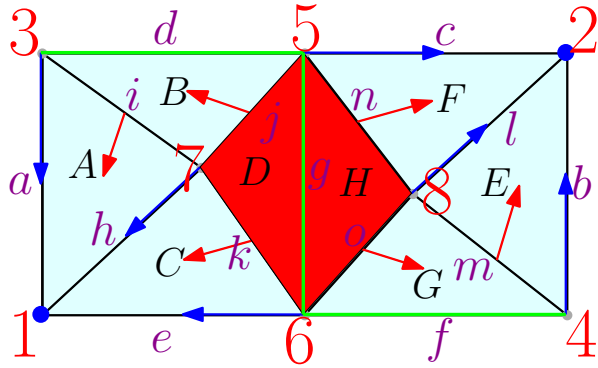
Figure 8 illustrates the result of applying this algorithm over the simple discrete gradient of figure 7 (note that the corresponding discrete function and gradient has been reproduced on figure 8(a)). The four diagrams displayed on figures 8(c) and 8(d) show the result obtained while computing the discrete extended ascending (left) and descending (right) 1-manifolds of the three saddle points (pink, yellow, and blue dashed lines), and the discrete extended ascending (left) and

descending (right) 2-manifolds of the two minima and two maxima (pink and yellow shaded regions) respectively.

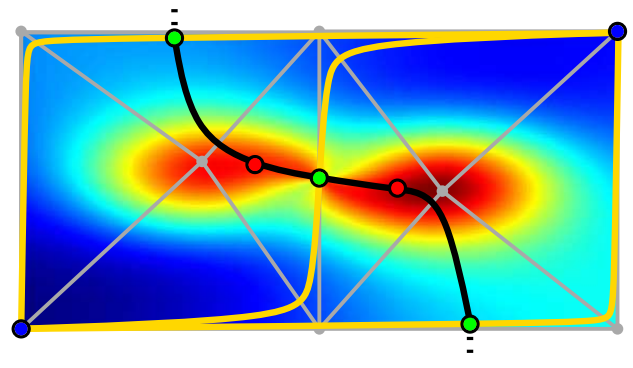
As an example, let us detail first the process followed by our algorithm to measure the ascending 1-manifold  $\mathcal{A}(C_1)$  of  $C_1$ , the critical 1-simplex (i.e. saddle point) with label  $d$  (see red path on left diagram of figure 8(c)). We start by considering the cofacets of  $C_1 = d$  and as there is only one, labeled  $B$ , we initially set  $A_{cur} = [B]$ . The 2-simplex  $B$  is linked to segment  $j$  by a gradient arrow,  $j$  is therefore added to  $\mathcal{A}(C_1)$  and  $A_{tmp} = j$ . Segment  $j$  has two cofacets, the triangles  $B$  and  $D$  and we therefore set  $A_{cur} = [B, D]$ . We then proceed by considering all triangles in  $A_{cur}$  one by one. The 2-simplex  $B$  is not critical but is paired to segment  $j$  which already belongs to  $\mathcal{A}(C_1)$ , it is therefore skipped and we are left considering triangle  $D$  which is critical and is therefore also skipped. Eventually, we obtain  $\mathcal{A}(C_1) = [j]$ . The pink path on the figure corresponds to the extended version of  $\mathcal{A}(C_1)$ , obtained by recursively also including the cofaces of the simplexes in  $\mathcal{A}(C_1)$ , namely the triangles  $B$  and  $D$ .

Similarly, the algorithm can be applied to the critical vertex  $C_0$  with value 1 to retrieve its ascending 2-manifolds displayed in pink on the left frame of figure 8(d). The cofacets of vertex 1 are segments  $a$ ,  $h$  and  $e$  and, as none of them is critical, the algorithm starts with  $A_{cur} = [a, h, e]$ . The segments in  $A_{cur}$  are paired with vertex 3, 7 and 6 respectively, which are not critical vertexes and do not yet belong to  $\mathcal{A}(C_0)$ , they are therefore added to  $\mathcal{A}(C_0)$  so that  $\mathcal{A}(C_0) = A_{tmp} = [3, 7, 6]$ . The content of  $A_{cur}$  is then replaced by all the segments that are cofaces of at least one vertex in  $A_{tmp}$ , and we have  $A_{cur} = [a, i, d, h, j, k, e, g, f]$ . Considering the segments in  $A_{cur}$  one by one,  $a$ ,  $h$  and  $e$  are skipped because they are paired to vertex 3, 7 and 6 respectively, which belong to  $\mathcal{A}(C_0)$ ,  $d$ ,  $g$  and  $f$  are skipped because they are critical and segments  $i$ ,  $j$  and  $k$  are skipped because they are not paired to 1-simplexes, but to the 2-simplexes  $A$ ,  $B$  and  $C$  respectively. This leaves  $A_{tmp}$  empty, and as a consequence  $A_{cur}$  becomes void which stops the algorithm with  $\mathcal{A}(C_0) = A_{tmp} = 3, 7, 6$ . The pink shaded region on the figure corresponds to the extended version of  $\mathcal{A}(C_0)$ , obtained by also adding the cofaces of vertex 3, 7 and 6, which are segments  $a$ ,  $h$ ,  $e$ ,  $d$ ,  $i$ ,  $j$ ,  $g$ ,  $k$ ,  $o$  and  $f$  and triangles  $A$ ,  $B$ ,  $C$ ,  $D$ ,  $G$  and  $H$ , as well as the extended ascending 1-manifolds of critical 1-simplexes  $d$ ,  $g$  and  $f$ .

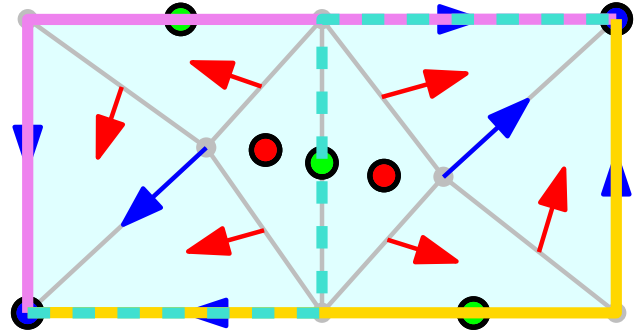
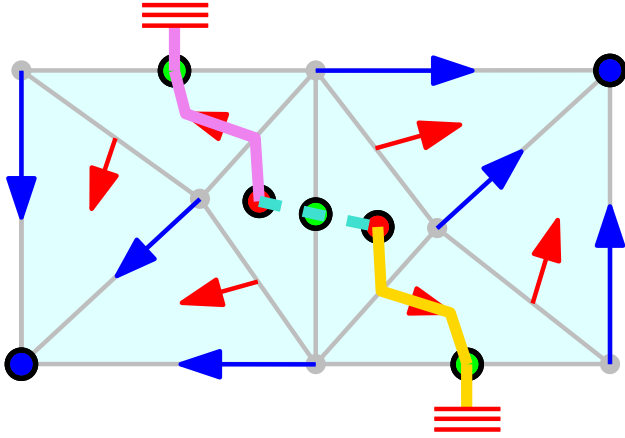
The computation of the arcs in the Morse-Smale complex is slightly more involved. A Morse-Smale complex is formed by critical nodes and arcs linking them together. Those arcs are integral lines that start at a critical point of order  $k+1$  and end at a critical point of order  $k$ , so they always have dimension 1: they are represented by curves. Their discrete equivalents are V-paths linking critical  $(k+1)$ -simplexes and critical  $k$ -simplexes. In 2D, they are simply described by the ascending and descending 1-manifolds (the dashed blue, pink and yellow lines on the upper part of figure 8(b)), but this is not the case in higher dimensions where arcs are generally described by the one dimensional intersections of a descending and an ascending manifold. The bottom diagram of figure 8(b) shows the discrete Morse Smale Complex computed from the simple density field  $\rho$  represented by the background



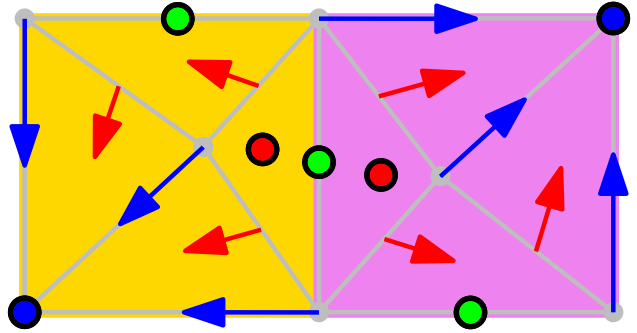
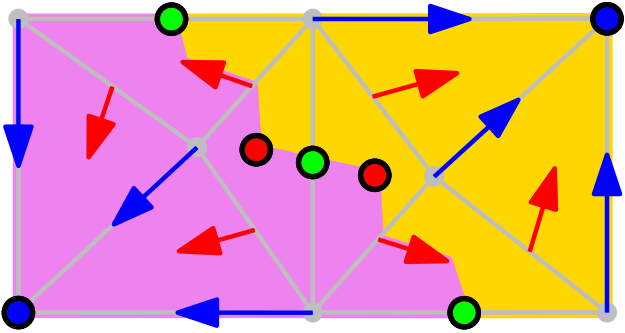
(a) Reproduction of the example discrete Morse function and its discrete gradient (see figure 7)



(b) Resulting discrete Morse-smale complex



(c) Computation of the discrete extended ascending (left) and descending (right) 1-manifolds (see definition 3.9) of the three critical 1-simplices (*i.e.* saddle points, the green disks, critical simplices being represented by disks for clarity)



(d) Computation of the discrete extended ascending (left) and descending (right) 2-manifolds (see definition 3.9) of the two critical 0-simplices (*i.e.* minima, the blue disks, critical simplices being represented by disks for clarity)

**Figure 8.** Illustration of the computation of the discrete extended ascending and descending manifolds and corresponding Morse-Smale complex from a discrete gradient. The application of the algorithm described in section 5.2 over the simple discrete function and gradient shown on panel 8(a) (see figure 7 for labels description) is illustrated on panels 8(c) and 8(d) for the ascending (left) and descending(right) 1-manifolds and 2-manifolds respectively (critical simplices are identified by colored disks in their center). On figure 8(c), the ascending (left diagram) and descending (right diagram) 1-manifolds of the three critical 1-simplices (*i.e.* equivalent of saddle points, represented by green disks) are represented as pink plain, cyan dashed and yellow plain broken lines respectively. The 1-manifolds are represented as sets of segments joining the center of simplices as according to definition 3.7, a V-path is an alternating sequences of  $k$  and  $(k+1)$ -simplices linked by a face/coface relation or belonging to a gradient pair. Note on the right diagram how it is possible for two descending 1-manifolds (blue dashed and plain yellow or plain pink) to share a portion of their path. On figure 8(d), the ascending (left diagram) and descending (right diagram) 2-manifolds of the two critical 0-simplices (*i.e.* equivalent of minima, represented by blue disks) are colored in pink and yellow respectively. The Morse-smale complex is the set of  $n$ -cell obtained by intersecting pairs ascending and descending manifolds (see definitions 2.7 and 2.8), and it is represented over the initial smooth function on panel 8(b). On this figure, the black and yellow curves represent the arcs (*i.e.* 1-cells) linking maxima/saddle points and minima/saddle points respectively. It is very striking how the algorithm manages to correctly capture the essential features of the Morse Smale complex, even though it was only applied over a very crude simplicial tessellation of space: not only the critical points were correctly identified as critical simplices, but the way they are connected by arcs is also correct (note that the arcs geometry was smoothed for clarity reasons).

color. It was obtained thanks to a modification of the manifold algorithm: when computing an ascending (resp. descending) manifold of a critical  $k$ -simplex, we store the list of critical  $(k+1)$ -simplexes (resp.  $(k-1)$ -simplexes) that are encountered and for each of them, trace the V-paths that led to them by storing in separated arrays all the simplexes in each path when the recursive procedure is returning. This way we obtain, for each pairs of critical  $k$ -simplex and  $(k+1)$ -simplex that are linked by a V-path, the set of all simplexes in the V-path (*i.e.* the arcs of the Morse-Complex). Note that on figure 8(b), each ascending (resp. descending) 1-manifold is actually made of two arcs, each linking the same saddle point to a maximum (resp. minimum). Algorithm 1 presents the pseudo code for a function that computes the ascending or descending arcs and manifolds of a critical  $k$ -simplex, the manifold and arcs being returned in global simplex array  $M$  and global list of simplex arrays  $arcs$  respectively. In this code, the lines dedicated to identifying arcs are tagged to differentiate them from the simpler manifold identification algorithm. After this function is called on a critical simplex  $C_k$ ,  $M$  will contain the index of all the  $k$ -simplexes in the ascending (resp. descending) manifold of  $C_k$  (not including  $C_k$ ) and  $Arcs$  will contain a list of arrays, each containing the  $k$ -simplexes in a V-path between  $C_k$  and another critical simplex  $C_{k+1}$  (resp.  $C_{k-1}$ ), including  $C_{k+1}$  (resp.  $C_{k-1}$ ) and  $C_k$ .

We end this subsection with a comment on our implementation. The ascending  $(d-k)$ -manifold and descending  $k$ -manifold of a critical  $k$ -simplex are represented by lists of  $k$ -simplexes. This certainly makes sense for the descending  $k$ -manifold, as in 3D for instance, volumes will be represented by lists of tetrahedrons, surfaces by list of faces and lines by lists of segments. However, this is not the case for the ascending  $(d-k)$ -manifold, where for instance, the ascending 3-manifold of a minimum is represented by a list of vertice. To solve this issue one can choose to use extended manifolds instead of regular manifolds. This can be problematic though, for instance for visualization purpose, not only because it considerably increases the number of simplexes within each manifold, but also because in that case two neighboring  $k$ -manifolds will edge  $k$ -simplexes. For instance, on figure 8(d), the extended ascending 2-manifolds should actually share 2-simplexes  $B$ ,  $D$ ,  $H$  and  $G$  if our algorithm was followed. It is not the case on the figure though because we actually used the dual tessellation for the representation of the extended ascending manifolds (*i.e.* the Voronoi tessellation in our case, where the complex is computed on a Delaunay tessellation). In fact, the dual tessellation associates a cell of dimension  $(d-k)$  to each  $k$ -simplex, and one only has to interpret the list of simplexes in an ascending manifolds in terms of its dual Voronoi cells, surfaces, lines or vertice, which does not necessitate any modification of the algorithm. Note that this point of view is also interesting as it enforces the fact that ascending and descending manifolds intersect transversely (*i.e.* they cannot be tangent in any point), an essential property of a Morse-Smale function (see section 2). In practice, we always use the dual representation for visualization of the descending manifolds,  $k$  dimensional regions being best represented by lists of  $k$ -simplexes, but

**Algorithm 1** Computes the ascending or descending manifold and arcs of a critical simplex  $S_k$ . Variables  $arcs$  and  $M$  store the retrieved Arcs and manifold. Triangular marks tag the lines dedicated to arcs identification only.

---

```

1: function GETMANIFOLD( $\sigma_k$ , ascending)
Require:  $\sigma_k$  is a critical  $k$ -simplex
Require:  $M$  is an empty list of simplexes
Require:  $arcs$  is an empty list of arrays of simplexes
2:   if ascending then
3:      $A_{cur} \leftarrow \text{GETCOFACES}(\sigma_k)$ 
4:   else
5:      $A_{cur} \leftarrow \text{GETFACES}(\sigma_k)$ 
6:   end if
7:    $A_{tmp} \leftarrow \{\}$ 
8:    $curArcs \leftarrow \{\}$ 
9:   for all  $c \leftarrow A_{cur}$  do
10:     $p \leftarrow \text{GETGRADIENTPAIR}(c)$ 
11:    if  $\text{GETDIMENSION}(p) == k$  and  $p \notin M$  then
12:       $M \rightarrow \text{INSERT}(p)$ 
13:      if not  $\text{ISCritical}(p)$  then
14:         $A_{tmp} \rightarrow \text{INSERT}(p)$ 
15:      else
16:         $newArc \leftarrow \{c\}$ 
17:         $arcs \rightarrow \text{PUSHBACK}(\&newArc)$ 
18:         $curArcs \rightarrow \text{INSERT}(\&newArc)$ 
19:      end if
20:    end if
21:  end for
22:  for all  $c \leftarrow A_{tmp}$  do
23:     $newArcs \leftarrow \text{GETMANIFOLD}(c, \text{ascending})$ 
24:     $curArcs \rightarrow \text{INSERT}(newArcs)$ 
25:  end for
26:  for all  $c \leftarrow curArcs$  do
27:     $c \rightarrow \text{PUSHBACK}(\sigma_k)$ 
28:  end for
29:  return  $curArcs$ 
30: end function

```

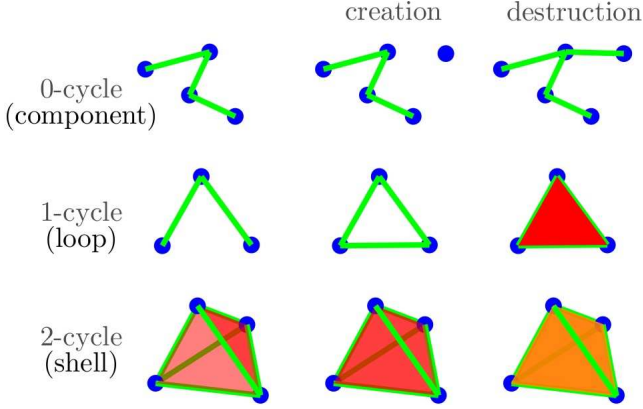
---

only store  $k$ -manifolds as lists of  $k$ -simplexes as this is much more efficient.

## 6 DEALING WITH NOISE: PERSISTENCE AND TOPOLOGICAL SIMPLIFICATION

Using the algorithms introduced in the previous two subsections, it is possible to compute efficiently the discrete Morse-Smale complex (DMC) of basically any function discretely sampled function via the delaunay tessellation of the sampling points. Applied directly to the Delaunay tessellation of a discrete galaxy catalogue or of a N-body dark matter simulation, the algorithm could therefore theoretically be used to identify the filaments, walls and void. However, because it cannot discriminate between the spurious Poisson noise induced detections and the actual cosmic web features, it is of no practical interest as is. As an example, we applied it to the output of a  $50 h^{-1}$  Mpc large dark matter simulation down-sampled to  $64^3$  particles. Running a simple friend-of-friend algorithm (Huchra & Geller 1982) with a linking length equal to one twentieth of the inter particular distance and a minimal number of particle of 20 leads to the identification of 800 bound structures (*i.e.* potential dark





**Figure 9.** Creation and destruction of  $k$ -cycles in a filtration according to  $F_\rho$  (equation 2). An unlinked component creates a 0-cycle, a loop around a hole creates a 1-cycle and a shell around an empty volume creates a 2-cycle.

matter haloes). Computing the Morse complex of the same distribution leads to the identification of 12771 maxima (*i.e.* potential haloes) and 32457 type 1 saddle points (*i.e.* potential filaments). This suggests that only about  $\sim 6\%$  of the detected structures are cosmologically significant and that most of the detected filaments actually link spurious noise induced maxima. In order to remedy this problem, we apply the concept of persistence (Edelsbrunner et al. 2000), introduced in section 4. Roughly speaking, persistence defines a mathematically rigorous framework to assess the significance of topological features while Morse theory, by the mean of the Morse-Smale complex, establishes the link between local geometry and topology. We describe in the following how, using those theories together, the Morse complex can be simplified in order to get rid of its unwanted features.

### 6.1 Pairing critical simplexes and persistence

Within the context of a smooth function, persistence can be understood as a measure of the life-time of a given topological feature (interpreted as the relative importance or significance of that feature) within the evolving sub-levels sets at levels varying from one extreme of the function possible values to the other. Within a discrete context, a very similar concept and interpretation can be defined for a filtration (see definition 4.2) of a simplicial complex  $K$ : new simplexes entering the filtration create or destroy topological features, defining their persistence in terms of how many new simplexes had to enter the filtration before a given topological feature was destroyed. In that case, one therefore measures the importance of the different topological features induced by the function that defines the order of entrance of each simplex in the filtration. As we are interested in the topology and geometry of the density function  $\rho$ , it is therefore natural to use its discrete counterpart  $F_\rho$  (see equation 2) to define the time each simplex enters the filtration, as it associate a distinct value to each simplex. We therefore consider the filtration  $F$  of  $K$

according to the ascending values<sup>8</sup> of  $F_\rho$ , similarly to what was done in section 5.2 to compute the Morse complex, and recast the persistence measure in terms of the difference of the value  $F_\rho$  associated to the simplex that creates a feature, and the simplex that destroys it.

Because of the way  $F_\rho$  was defined, any simplex enters  $F$  before its cofaces. In the 3D case for instance, this is illustrated on figure 9. A vertex (0-simplex) is never linked to the rest of  $F$  when it enters and therefore we say that it will always *create* a new component (*i.e.* a 0-cycle) in the filtration. Similarly, when a segment enters, its two faces already belong to  $F$  while its cofaces do not yet: it forms a bridge between two 0-simplexes, and may therefore either *destroy* one component if those two 0-simplexes belonged to distinct “islands” or *create* a new 1-cycle (*i.e.* a loop, a torus like structure) in the other case. The same way, a face could *destroy* a 1-cycle by filling the hole in its center or *create* a 2-cycle (*i.e.* a shell), and a tetrahedron may only destroy a 2-cycle (*i.e.* fill a shell). A consequence of the fact that all simplexes create or destroy something is that all simplexes in the complex are initially critical, as was already noted in section 5.1, and our goal is to establish which critical simplexes respectively create and destroy a given  $k$ -cycle of  $F$  (*i.e.* a component, a loop, a shell, ...).

Actually, that is exactly what the algorithm that computes the discrete gradient does for the so called  $\epsilon$ -persistent arcs (*i.e.* arcs that link simplexes whose value  $F_\rho$  only differs by an infinitesimal amount  $\epsilon$ , see section 5.1). In fact, a simplex  $\sigma_k$  and its face  $\sigma_{k-1}$  may belong to a gradient pair if their value differ only by  $\epsilon$  (*i.e.* if they enter consecutively in the filtration). When this is the case, the value of  $F_\rho$  is symbolically modified by an infinitesimal amount so that  $\sigma_{k-1}$  actually enters just before  $\sigma_k$  and none of them may create or destroy a cycle anymore. Whereas  $\sigma_k$  created a new  $k$ -cycle destroyed by  $\sigma_{k-1}$  in the initial filtration, this is not the case anymore after the modification, both simplexes are not critical anymore, and belong to gradient pair instead. We therefore only need to pair the critical simplexes that survive to the discrete gradient computation (*i.e.* that belong to the discrete Morse-Smale complex) into persistence pairs. Edelsbrunner et al. (2000) first introduced an algorithm that does exactly that in 3D. Although, more general and efficient approaches have been developed since (*e.g.* Cohen-Steiner et al. (2006) or Zomorodian (2009)), we present here a variation of the original one, directly implemented over the morse complex. Note that given that only the critical simplexes identified in the DMC of  $K$  create or destroy cycles, one only needs to consider the Morse complex directly (*i.e.* as opposed to considering each and every simplex in  $K$ ) to identify persistence pairs. From that point, it therefore does not matter anymore how the Morse-Smale complex was computed, or whether it is discrete or not, as both have identical combinatorial properties anyway. Under the assumption that the discrete Morse function was computed with enough care to correctly inherit the topology of the underlying density field, we can

<sup>8</sup> note that the choice of ordering according to ascending or descending value is totally arbitrary and has no importance.

therefore indifferently talk about the critical points of the smooth density field  $\rho$  or the critical  $k$ -simplexes  $\sigma_k$  of the simplicial complex  $K$  in the following. It is also equivalent to describe persistence in term of creation/destruction events in the level-sets of  $\rho$  or in the filtration steps of the filtration induced by  $F_\rho$  (by convention, we choose to order the entrance time by ascending values of  $F_\rho$ ).

The algorithm starts by tagging each critical simplex  $\sigma_k$  as *positive* or *negative* depending on whether it creates or destroys a cycle. As was noted before, in 3D, the critical vertices and tetrahedrons (equivalent of minima and maxima) may only create a 0-cycle and destroy a 2-cycle respectively. The critical 0-simplexes are therefore all tagged positive and the critical 3-simplexes are negative. The sign of the rest of the critical simplexes is determined by following the growth and merging of each component in the filtration using a “union-find” type data structure<sup>9</sup>. Depending on whether a segment entering the filtration links two previously independant components (*i.e.* destroys a 0-cycle) or creates a new bridge within one unique component (*i.e.* creates a 1-cycle), it will be tagged negative or positive as it destroys a component or creates a 1-cycle. Tracking the creation of 2-cycles (*i.e.* shells) or destruction of 1-cycles by the critical 2-simplexes in the filtration seems much more complex though, but it can actually be made easy by considering the filtration  $F'$  induced by  $-F_\rho$ , where simplexes enter in the opposite order to  $F$ . For symmetry reasons, a 2-simplex creating a 2-cycle in  $F$  actually destroys a component in  $F'$ , and is therefore positive, while a simplex destroying a 1-cycle in  $F$  actually creates 1-cycle in  $F'$  and is therefore negative. The exact same algorithm and “union-find” type data structure can therefore be used to track those events in  $F'$  and decide the sign of each critical 2-simplex in  $F$ .

Practically, let us consider the filtration in the ascending order first. An entry is created in a “union-find” structure for each critical simplex in the DMC, each of them is initially attributed a different group Id. Whenever a segment enters the filtration, the group Id of its two facets are retrieved and we check if they differ or are equal. In the first case, this means the segment created a bridge between two previously disjoint structures. It is therefore tagged negative and the groups of the two vertices and the segment are merged in the union find structure. In the second case, both vertice already belonged to the same structure, which means that the introduction of the segment created a new 1-cycle (*i.e.* a loop that passes through the newly created bridge). The segment is therefore tagged positive and its group is merged with that of its faces. The sign of the 2-simplexes (triangles) is determined in the same way, but reversing the order of the filtration: a face is tagged positive whenever it creates a bridge between two previously unlinked tetrahedron and negative

whenever it links two tetrahedron that where already linked.

---

**Algorithm 2** Finds persistence cycles created by a negative critical simplex  $\sigma_k$ .

---

```

1: function CYCLESEARCH( $\sigma_k$ )
Require:  $\sigma_k$  is a negative critical  $k$ -simplex (parameter)
Require:  $ppairs$  stores persistence pairs (global)
Require:  $cycles$  store all previously computed cycles, each
        associated to a negative simplex (global)
Require:  $CurSet$  is a  $\mathbb{Z}_2$ -Set of simplexes (see text), empty
        when first called (local)
2:    $\alpha_{nei} \leftarrow \text{GETMSCNEIGHBORS}(\sigma_k)$   $\triangleright \alpha$  contains the
        simplexes that share an arc with  $\sigma_k$  in the DMC.
3:   for all  $\beta \leftarrow \alpha_{nei}$  do
4:     if  $\text{TYPEOF}(\beta) == \text{TYPEOF}(\sigma_k) - 1$  and not
         $\text{SIGNOF}(\beta) == \text{SIGNOF}(\sigma_k)$  then
5:        $CurSet \rightarrow \text{INSERT}(\beta)$ 
6:     end if
7:   end for
8:   while not  $\text{ISEMPTY}(CurSet)$  do
9:      $\sigma_{k-1}^{cur} \leftarrow \text{GETHIGHESTOF}(CurSet)$ 
10:    if  $\text{ISEMPTY}(cycles[\sigma_{k-1}^{cur}])$  then
11:       $cycles[\sigma_{k-1}^{cur}] \leftarrow CurSet$ 
12:       $cycles[\sigma_k] \leftarrow CurSet$ 
13:       $ppairs \rightarrow \text{INSERT}(\sigma_{k-1}^{cur}, \sigma_k)$ 
14:      break
15:    else
16:      for all  $\beta \leftarrow cycles[\sigma_{k-1}^{cur}]$  do
17:         $CurSet \rightarrow \text{INSERT}(\beta)$   $\triangleright$ 
        note that adding the cycle of  $\sigma_{k-1}^{cur}$  modulo 2 actually
        removes simplex  $\sigma_{k-1}^{cur}$ .
18:      end for
19:    end if
20:  end while
21: end function

```

---

Now that each critical simplex  $\sigma_k$  has been attributed a sign, we can reconsider the filtration  $F$  of the critical simplexes in ascending order, and identify the persistence pairs using algorithm 2. Instead of detailing how this rather complex algorithm works, let us detail its application to a simple 2D example for the sake of clarity. Note that the method is very similar whatever the number of dimensions, as long as the sign of each critical simplex has been previously determined, and so deducing the 3D case from the 2D one should be straightforward. We first define a few variable names and types the algorithm uses. The purpose of the function  $cycleSearch(\sigma_k)$  is to retrieve the  $(k-1)$ -cycle destroyed by the negative critical simplex  $\sigma_k$ . For each call, the result is stored in a variable  $cycles$  that will in the end contain the description of all cycles, each associated to its creating and destroying critical  $k$ -simplex. Each cycle is stored as a list of critical  $(k-1)$ -simplexes that form a  $(k-1)$ -cycle within the Morse-Smale complex (for instance, a loop is stored as a list of critical segments). Another variable, labeled  $ppairs$ , stores pairs of critical simplexes that creates or destroys its corresponding a given cycle  $[\sigma_k, \sigma_{k-1}]$ . Basically, the function  $cycleSearch(\sigma_k)$  is called once every time a negative critical simplex  $\sigma_k$  enters the filtration. Internally, the function uses a variable

<sup>9</sup> a Union-find data structure is particularly efficient at managing large number of sets of elements. It implements fast set merging (the “union” operation) and is able to recover efficiently to which set a given element belongs to (“find” operation).

*CurSet*, of special type “ $\mathbb{Z}_2$ -Set”, to store a temporary list of critical  $(k-1)$ -simplexes considered at a given time. The type “ $\mathbb{Z}_2$ -Set” implements the  $k$ -chain group addition of definition B.1, or in other words it behaves like regular “Set” structure, that stores sets of elements, but contrary to normal “Set”, adding an element already contained in the  $\mathbb{Z}_2$ -Set results in its actual removal<sup>10</sup>.

We show on figure 10 the aforementioned practical example of persistence pairs and corresponding  $k$ -cycles computation over a simple Morse-Smale complex. The upper left frame shows a DMC computed from a high resolution triangulation of the underlying density field  $\rho$  (note that only the smooth function is represented, not the simplicial complex). As mentioned earlier, only the structure of the Morse-Smale complex is necessary to identify the cycles, so on these figures, we represented in the background the sub-level sets of the density field  $\rho$  instead of steps in the filtration  $F_\rho$  to show how cycles are created and destroyed. We could identically have shown subsets of a simplicial complex, and actually, at this stage, we could equally say that the colored disks represent minima/critical points/maxima of the smooth field  $\rho$  or critical vertexes/segments/triangles (*i.e.* 0/1/2-simplexes) of the discrete Morse function  $F_\rho$ .

A selection of 12 steps corresponding to the entrance in the ascending filtration of 12 of the 21 critical simplexes are represented on the frames in the bottom part. The entrance of the first 8 critical simplexes (blue disks) is not represented and the first displayed step, step 9, corresponds to the entrance of the rightmost critical 1-simplex (green disk). Note however that before step 9, the critical vertexes (*i.e.* minima) from 1 to 7 already entered creating each one component in the filtration, and critical segment 8 also entered, destroying the 0-cycle created by critical vertex 7 which was merged with that of vertex 2 (this destruction is still represented at step 9 by the pink and red lines though). Considering critical segment  $\sigma_1 = 9$ , we first retrieve its two neighboring critical 0-simplexes, labeled 4 and 3, and we therefore have  $CurSet = \{3, 4\}$ . We first consider the highest,  $\sigma_0^{cur} = 4$ , and check if there is a cycle associated to it in  $cycles[\sigma_0^{cur}]$ . As this is not the case, it means that we have found the cycle of  $\sigma_1$ , and therefore set  $cycles[\sigma_1] = cycles[\sigma_0^{cur}] = CurSet = \{3, 4\}$  and insert pair  $[\sigma_1, \sigma_0^{cur}] = [9, 4]$ . On panels 9, all the critical simplexes involved in the cycle are circled in pink, the pink arc connects the critical simplexes in the identified pair and the red line represents the cycle. For instance, in that case, we identified that critical segment 9 destroys the component (0-cycle) created by critical vertex 4, which results in the components created by critical vertexes 3 and 4 merging into each other (see the sub-level sets in the background).

Step 10 is very similar to step 9, with critical segment  $\sigma_1 = 10$  entering, and we therefore add persistence pair  $[10, 6]$  to  $ppairs$  and set  $cycles[\sigma_1] = cycles[\sigma_0^{cur} = 6] = \{6, 2\}$ . Step 11 is skipped

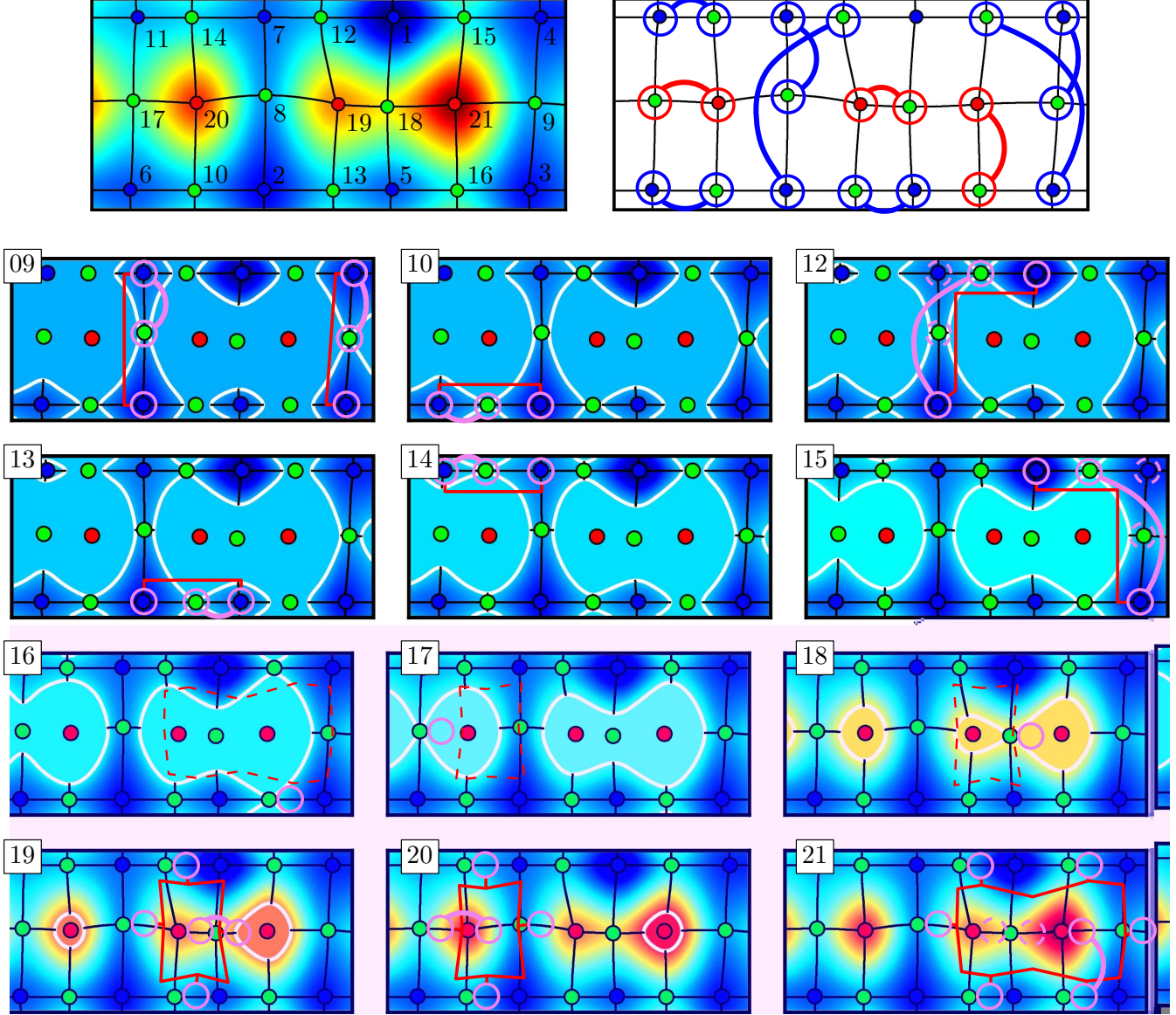
as it corresponds to the entrance of a positive critical vertex (*i.e.* the creation of a new component), but step 12 is more interesting. Critical segment 12 is negative, and we therefore start the algorithm as previously by setting  $CurSet = \{7, 1\}$ , its neighbor critical vertex on the DMC. The highest critical vertex in  $CurSet$  is  $\sigma_0^{cur} = 7$ , which was already paired at step 8 (represented on panel 9). We therefore add the cycle associated to it,  $cycles[\sigma_0^{cur} = 7] = \{2, 7\}$ , to  $CurSet$ , which gives  $CurSet = \{7, 1\} + \{2, 7\} = \{1, 2\}$ , as the addition is performed modulo 2 ( $CurSet$  is of type  $\mathbb{Z}_2$ -Set). The new highest critical vertex in  $CurSet$  is therefore  $\sigma_0^{cur} = 2$ , which is not paired yet. We therefore add the new pair  $[12, 2]$  to  $ppairs$ , and set  $cycles[\sigma_1 = 12] = cycles[\sigma_0^{cur} = 2] = CurSet = \{1, 2\}$ , which basically means that when critical segment 8 enters, the component created by vertex 2 merges into that of vertex 1. Steps 13 and 14 correspond to simple pairings (similar to step 9), and step 15 is similar to step 12, as critical vertex 4 is already paired, resulting in variables  $ppairs$  and  $cycles$  being updated according to the diagrams of panel 13, 14, and 15.

The critical segment entering at step 16 is different though, as it creates a 1-cycle (symbolized by red dashed lines on panel 16). Indeed, its neighbors critical vertexes on the DMC are 3 and 5, which already belong to the same component at step 16 (as can be seen on the underlying sub-level set or on the DMC, by observing that the path  $[5, 13, 2, 8, 7, 12, 1, 15, 4, 9, 3]$  only has critical simplexes with values below 16). As this critical segment is therefore positive, we skip it for now, but we will see later how its cycle will be identified when it gets destroyed by a critical 2-simplex. The following steps 17 and 18 are similar, and corresponding critical segments are skipped.

The first negative critical 2-simplex enters at step 19. Following algorithm 2, we start with  $CurSet = \{8, 12, 18, 13\}$ , the four critical segments neighbors of the critical triangle 19 on the DMC. The highest valued critical vertex in  $CurSet$  is  $\sigma_1^{cur} = 18$ , which is not yet paired, and we therefore add pair  $[19, 18]$  to  $ppairs$  and set  $cycles[\sigma_2 = 19] = cycles[\sigma_1^{cur} = 18] = \{8, 12, 18, 13\}$ , represented by the red loop on panel 19 (see also red dashed loop of figure 18, when the cycle was created). This means that critical segment 18 created a new 1-cycle that was destroyed by critical triangle 19, and this cycle is represented by the closed path formed by critical segments  $\{8, 12, 18, 13\}$ , which are linked to each other through their neighbor critical vertexes in the DMC, 1, 7, 2 and 5 (the 1-cycle is given by sequence  $[18, 5, 13, 2, 8, 7, 12, 1, 18]$ , which can be easily retrieved at query time from the information in  $ccycles$ ). Critical triangle 20 also destroys a 1-cycle. The process is similar to previous step and variables are updates accordingly. We finally proceed to step 21, where the last critical simplex enters. It is also negative (as all critical triangles are anyway), and we start with  $CurSet = \{9, 15, 16, 18\}$ . The highest critical segment is  $\sigma_1^{cur} = 18$ , which is already paired to critical triangle 19, and its cycles is therefore added modulo 2 to  $CurSet$ , giving  $CurSet = \{9, 15, 16, 18\} + cycles[\sigma_1^{cur} = 18] = \{9, 15, 16, 18\} + \{8, 12, 18, 13\} = \{9, 15, 16, 8, 12, 13\}$ . As critical segment 16, the highest in  $CurSet$ , is free, this

<sup>10</sup> hence the name, as each element behaves as if it was counted modulo 2, with coefficients in  $\mathbb{Z}_2$





**Figure 10.** Illustration of the computation of persistence pairs using the algorithm 2 presented in section 6.1. On the top left frame, the Morse-smale complex of the underlying smooth function  $\rho$  is represented with blue, green and red disks, standing for minima, saddle points and maxima. Note that the Morse complex is actually a DMC computed from a discrete morse function  $F_\rho$  (see equation 2) over a high resolution tessellation (not represented), so the blue, green and red disks, equally stand for critical vertexes, segments and triangles respectively (the two views are equivalent under the assumption that  $F_\rho$  correctly identify the topology of  $\rho$ ). The numbers  $n$  beside the disk correspond to the corresponding values of the density  $\rho$ . On the 12 panels in the bottom part, the evolution in the sub-level sets (*i.e.* the set of points where density  $\rho$  is smaller than a given threshold  $\rho_n$ ) of the smooth density field is shown in the background, at levels  $\rho_n = n$  corresponding to the value  $n$  in the upper left corner of each panel. On each panel, the identification of a new persistence pair in the DMC is represented by a pink arc, while the corresponding cycle is symbolized by the red line. Note that cycles and pairs are identified at the moment they are destroyed, not created, and red dashed lines on panels 16, 17 and 18 roughly indicate the shape of the 1-cycles (*i.e.* 1D loop) at the moment of their creation, for information. The pink plain and dashed circles indicate all nodes of the DMC that are concerned by the creation or destruction of a cycle at a given step. Finally, all the identified persistence pairs are represented on the top right frame, in blue or red depending on their type. A detailed description of the computation of the persistence pairs and  $k$ -cycles as shown on this figure is given in the main text (see second half of section 6.1).

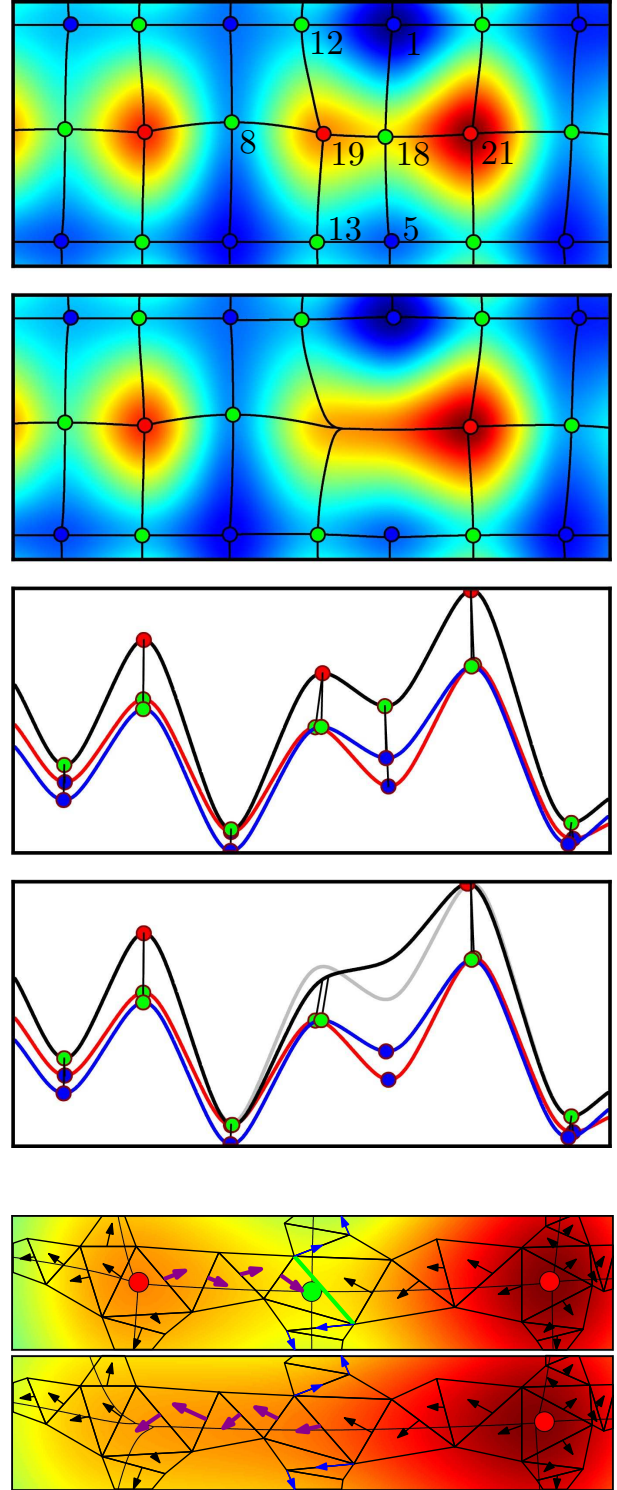
means we are done identifying the last cycle. We therefore update variables *ppairs* and *cycles* accordingly and the algorithm terminates.

The upper right frame shows all of the persistence pairs over the DMC, and one can convince himself of the correctness of the cycles retrieved at steps 19, 20 and 21 by comparing them to what they actually looked like in the sub-level sets when they were created, at steps 16, 17 and 18 (see the dashed red cycles and newly created closed loops in the white iso-contours in the background). Although we do not show examples here, the method is strictly similar in higher dimensions.

## 6.2 Simplification

The relative importance of topological features can be reliably assessed using persistence theory, and it was briefly shown in section 4 how it is possible in the 1D case to locally modify a smooth function in order to cancel a low persistence pair of critical points without affecting other critical points. Although it would also seem a viable option to directly modify  $\rho$  in order to cancel non persistent pairs in the higher dimensions, this may not be the best thing to do. From a purely technical point of view, for large data sets, the computational cost of actually modifying  $\rho$  and recomputing the Morse-Smale complex every time would be excessive. From a theoretical point of view, one would need to arbitrarily define a more or less natural way to smoothly transform  $\rho$  so that the canceling pairs would disappear without affecting the remaining critical points. Fortunately, such a transformation does not need to be explicitly conducted and it is enough to know that it exists and how it affects the Morse-Smale complex.

As a simple example, an arbitrary modification<sup>11</sup> of the smooth density field of figure 10 that cancels the low persistence pair [18,19] is presented on the top frames of figure 11. As expected, this modification of  $\rho$  leads to the removal of the saddle-point and maximum, and a particular reorganization of the arcs in the Morse Complex. Before cancellation, the saddle point 18 was linked to two minima (1 and 5) and two maxima (19 and 21). With its removal, the arcs emanating from the minima get rerouted to maximum 21 (as maximum 19 is also canceled in the operation), and they are therefore not critical anymore: they are removed from the Morse complex. The situation is different for saddle points 8, 12 and 13 though, which were linked to maximum 19. During the cancellation, the gradient of  $\rho$  is reversed between the canceled points (see lower panels), and the arcs that led to maximum 19 are therefore free to continue their ascension up to the former position of the canceled saddle point, and further along the arc [18,21], to finally reach maximum 21. Those field lines still link saddle-points to maxima, they are criti-



**Figure 11.** Topological simplification of a maximum and saddle point persistence pair in the smooth 2D field  $\rho$  of figure 10. On the upper part, from top to bottom, the four frames display the Morse complex before and after simplification and the corresponding density profiles along the three horizontal axis of the Morse complex (red, black and blue for the upper, middle and lower axis respectively). The density profile before simplification is represented in gray. The lowest frame shows an equivalent cancellation of critical pairs in a discrete Morse complex by discrete gradient reversal. Note that non essential gradient pairs and simplexes have been omitted for clarity, as they are not affected by the path reversal.

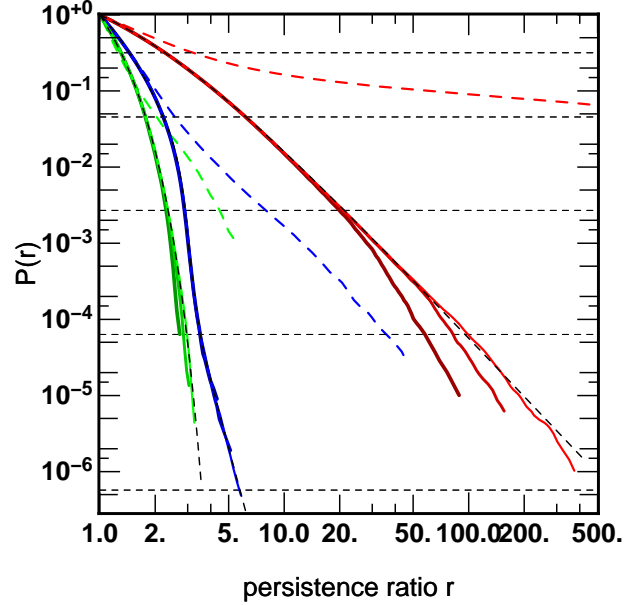
<sup>11</sup> Note that achieving the modification shown in this example was actually made easy by the fact that the function itself is analytically defined as a sum of Gaussian functions, but it would clearly have revealed much more challenging in the general case.

cal and therefore belong to the new modified Morse complex.

The field lines reorganization scheme during a cancellation can be intuitively understood in the general case by defining a similar minimalistic transformation of a discrete Morse function and its discrete gradient. Basically, the essential feature lies in the reversal of the gradient path between the canceled critical points. Such an operation can easily be defined over a discrete gradient (Forman 2002). The corresponding modification is shown on the bottom frame of figure 11, in the case of a discrete Morse function similar to  $\rho$  and defined over a tessellation: by reversing the path of discrete gradient arrows between the critical points (purple shade), the two critical points are effectively removed while the rest of the discrete gradient is left unmodified, and it is easy to predict the consequences of this modification on the discrete Morse complex. Let us call  $\sigma_k$  and  $\sigma_{k+1}$  the critical  $k$  and  $(k+1)$ -simplexes to cancel, and  $\alpha_{k+1}^i$  and  $\beta_k^j$  the critical  $k+1$  and  $k$ -simplexes respectively linked to  $\sigma_k$  and  $\sigma_{k+1}$  by an arc in the DMC. By reversing the gradient path between  $\sigma_k$  and  $\sigma_{k+1}$ , one also reroutes all the arcs and manifolds that previously reached one of those critical simplexes. After cancellation, an ascending arc emanating from  $\beta_k^j$  still reaches the formerly critical simplex  $\sigma_{k+1}$ , and it can be extended through the reversed path and continue following any previously ascending arc emanating from  $\sigma_k$ , leading to a critical  $(k+1)$ -simplex  $\alpha_{k+1}^i$ . Similarly, any descending  $(k+1)$ -manifold of  $\alpha_{k+1}^i$  now reaches  $\sigma_{k+1}$  and can therefore be extended by the descending  $(k+1)$ -manifold of  $\sigma_{k+1}$ . For the same reason, the ascending  $(d-k)$ -manifolds of  $\beta_k^j$  can be extended by the ascending  $(d-k)$  manifolds of  $\sigma_{k+1}$ . One therefore does not need to actually perform any gradient path reversal, and the cancellation of critical pair  $[\sigma_k, \sigma_{k+1}]$  is achieved directly on the DMC using the following procedure:

- (i) Let  $\alpha_{k+1}^i$  and  $\beta_k^j$  be the  $N_\alpha$  and  $N_\beta$  critical  $k+1$  and  $k$  critical simplexes sharing an arc in the DMC with  $\sigma_k$  and  $\sigma_{k+1}$  respectively.
- (ii) Create a new arc between each of the  $N_\alpha * N_\beta$  pair  $[\alpha_{k+1}^i, \beta_k^j]$  by joining arcs  $[\alpha_{k+1}^i, \sigma_k]$ ,  $[\sigma_k, \sigma_{k+1}]$ , and  $[\sigma_{k+1}, \beta_k^j]$ . The path  $[\sigma_k, \sigma_{k+1}]$  must be reversed during the operation.
- (iii) Extend the descending manifold of each  $\alpha_{k+1}^i$  with the descending manifold of  $\sigma_{k+1}$ .
- (iv) Extend the ascending manifold of each  $\beta_k^j$  with the ascending manifold of  $\sigma_k$ .
- (v) Delete the critical simplexes  $\sigma_k$  and  $\sigma_{k+1}$ , together with their 4 ascending and descending manifolds and all of the arcs leading to or emanating from them.

It is important to remark that in general, the simplification of a pair may lead to an increase in the total number of arcs in the complex. This is particularly true when none of the canceling simplex is a 1-saddle or a  $D-1$  saddle, as in that case, the number of arcs is not bounded. Moreover, there exist two specific cases where a cancellation is impossible. The first is when critical simplexes do not share an arc in the DMC. The second is when they share more than one arc, as in that case, gradient path reversal would lead to the creation of a looping path in the discrete gradient, which is forbidden (see section 3). The



**Figure 12.** The cumulative probability  $P_k(r)$  that a persistence pair of order  $k$  with persistence ratio greater or equal to  $r$  exists in a 3D scale free Gaussian random field (colored plain curves) and in a  $50h^{-1}$  Mpc dark matter cosmological simulation (colored dashed curves). The red, blue and green curves correspond to maxima/1-saddle, 1-saddle/2-saddle and 2-saddle/minima pairs respectively. The different shades, from darker to lighter, correspond  $64^3$ ,  $128^3$  and  $192^3$  particles realizations respectively. The black dashed curves show fits to the Gaussian case, as presented in the main text, while the horizontal dashed lines correspond to different significance levels in units of “sigma”, ranging from  $S = 1 - \sigma$  (top) to  $S = 5 - \sigma$  (bottom).

detailed procedure to deal with this is explained in section 7.

### 6.3 Filtering Poisson noise

As mentioned previously, mainly because of Poisson noise, it is not possible to use the raw DMC to identify structures in the cosmic web. In fact, most of the critical points, arcs and manifolds are actually spurious artifacts created by sampling noise. This is especially true in the present case, where we wish to use DTFE density and a simplicial complex computed from the Delaunay tessellation of a discrete realization. As a matter of fact, the scale free nature of DTFE makes it locally very sensitive to Poisson noise, as information is always locally extracted at the sample resolution limit. Our approach to remedy this problem consists in computing a significance level for each persistence pair, and canceling the persistence pairs whose significance is below a given threshold.

Let  $r$  be the persistence ratio of a persistence pair  $q_k = [\sigma_k, \sigma_{k+1}]$ , then

$$r(q_k) = F_\rho(\sigma_{k+1}) / F_\rho(\sigma_k). \quad (3)$$

We note  $P_k(r_0)$  the cumulative probability that a persistence pair of critical simplexes of order  $k$  and  $k+1$  and with persistence ratio  $r \geq r_0$  exists in the Delaunay



tessellation of a random discrete Poisson distribution. It is then convenient to denote the relative importance of a given critical pair  $q_k$  in terms of its significance,  $S(q_k)$ , expressed in units of “sigmas” with analogy to the Gaussian case:

$$S(q_k) = S_k(r(q_k)) = \text{Erf}^{-1}\left(\frac{P_k(r(q_k)) + 1}{2}\right), \quad (4)$$

where Erf is the error function. As a purely analytical derivation of  $P_k(r)$  seems clearly out of reach, we use Monte-Carlo simulation to estimate it, measuring  $P_k(r_0)$  as the fraction of persistence pairs of order  $k$  with persistence ratio  $r \geq r_0$  in a Poisson sample. The results are shown on figure 12. On that figure, the values of  $P_k(r)$  is plotted as a function of  $r$  in green, blue and red for  $k = 0$ ,  $k = 1$  and  $k = 2$  respectively and the horizontal dashed lines represent different significances level in units of “sigma”, ranging from  $S = 1-\sigma$  (top) to  $S = 5-\sigma$  (bottom). From these results, the following fits can be extracted in the 3D case (represented as black dashed lines on figure 12):

$$P_0(r) = \exp(-\alpha_0(r-1) - \alpha_1(r-1)^{\alpha_2}) \quad (5)$$

with  $\alpha \approx [3.694, 0.441, 2.538]$ ,

$$P_1(r) = f_1 \cdot (1-t) + f_2 \cdot t \quad (6)$$

with  $f_1 = \exp(-\beta_0(r-1))$ ,  $f_2 = \beta_1 r^{-\beta_2}$   
 $t = (1 + \beta_3/u^{\beta_4})^{-1}$   
 $\beta \approx [2.554, 4.000, 9.000, 1.785, 14.000]$ ,

$$P_2(r) = (1 + \gamma_0(r-1))^{-\gamma_1} \quad (7)$$

with  $\gamma \approx [0.449, 2.563]$ ,

and in the 2D case, we obtain:

$$S_0^{2D}(r) = \exp(-\alpha_0 \cdot (r-1) - \alpha_1 \cdot (r-1)^{\alpha_2}) \quad (8)$$

with  $\alpha \approx [2.00, 0.01, 3.50]$ ,

$$S_1^{2D}(r) = (r-1)^{-\beta_0(1+\beta_1 \log(r-1))} \quad (9)$$

with  $\beta \approx [0.75, 0.20]$ .

A relatively similar approach was undertaken in a code named ZOBOV Neyrinck (2008) to measure the significance of cosmological voids. The approach developed in this article nevertheless differs from ours in that it is limited to voids and that they do not use persistence pairs. Instead, they pair minima of the density field to the lowest 1-saddle point on the surface of their ascending 3-manifolds (*i.e.* the voids themselves) that is not already paired to another minimum with higher density. This explains why our fit of  $P_0(r)$  differs from theirs.

The fact that the expression of the fits for  $k = 0$  and  $k = 2$  is relatively simple compared to the one for  $k = 1$  may seem intriguing at first sight. But if the fit for function  $P_1(r)$  actually requires more coefficients, it is mainly because it undergoes some sort of transition around  $r = 4$ , which roughly corresponds to a significance level of  $3.5-\sigma$ . We believe that this only reflects the nature of DTFE itself, whose probability distribution function is clearly biased toward the high densities as the number of minima is limited

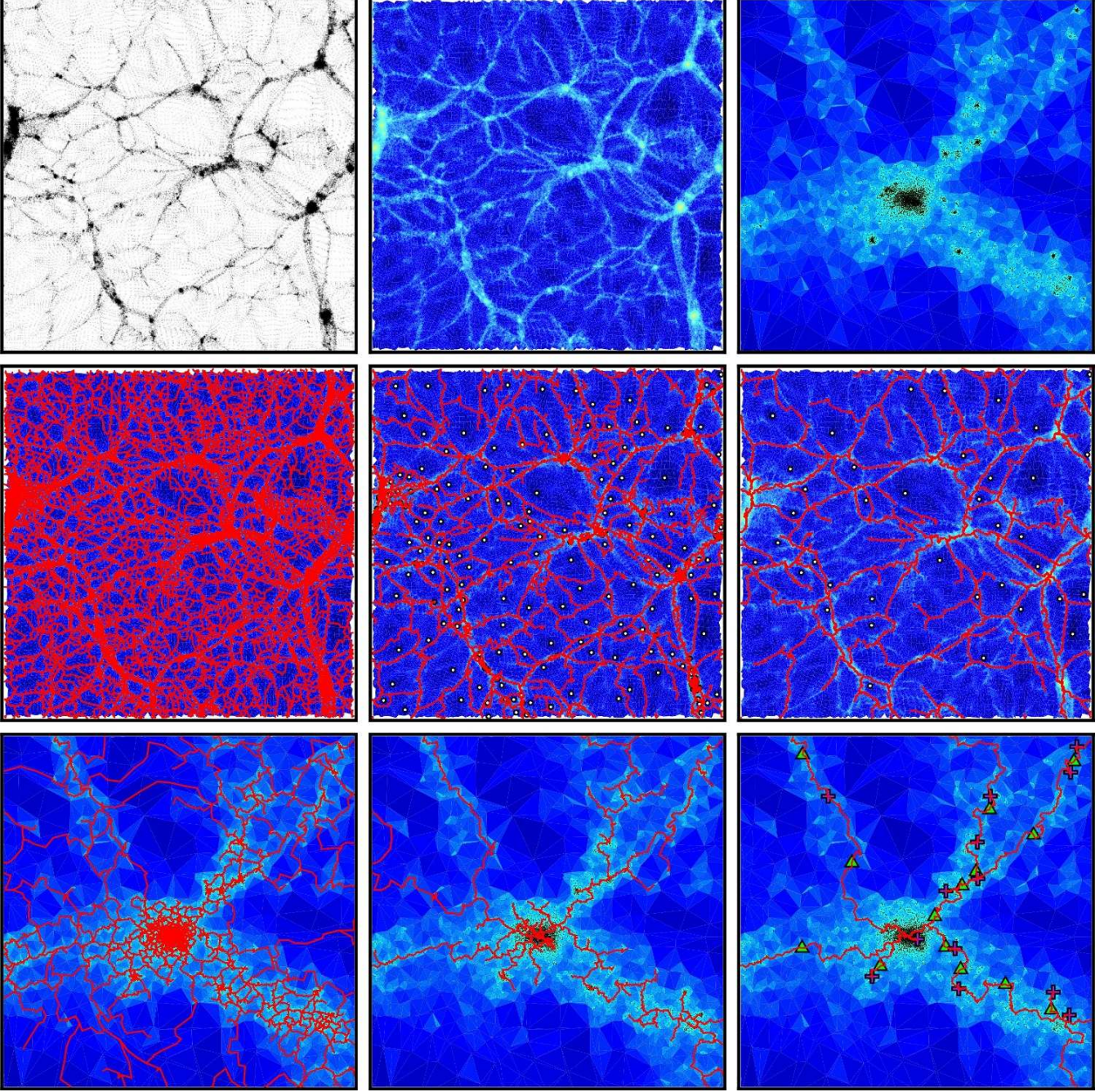
by the comparatively larger volume of the Voronoi cells they occupy (see Schaap & van de Weygaert (2000)). As the size of our Monte-Carlo sample is limited, just like this results in an increase of the number of  $k = 1$  type pairs when fewer and fewer comparatively lower density minima become available to form pairs. We also note that this tendency is present in the cumulative probabilities of persistence pairs ratio measured in cosmological simulations as well (colored dashed curves). Nevertheless, those probabilities are significantly higher than that in the Poisson sample for any value of  $k$  and it therefore seems that it should be reasonably easy to filter spurious persistence pairs without affecting too much those storing precious information on the cosmic web topology.

#### 6.4 Illustration in 2D

Figure 13 shows the DMC of a 2D discrete distribution of  $\sim 350,000$  particles with periodic boundary conditions computed at different levels of significance. The discrete distribution (upper left) was obtained by projecting a sub-sampled  $10 h^{-1}$  Mpc slice of a  $50 h^{-1}$  Mpc large dark matter N-Body simulation at redshift  $z = 0$ . The resulting delaunay tessellation, composed of  $\sim 1,000,000$  1-faces and 670,000 2-faces, and corresponding DTFE density field are shown on the upper central and upper right panels. Note that identifying the filamentary structure in such a distribution is particularly challenging because of its very high dynamic range and also because many filaments simply disappear into low density regions as they leave the slice in the original 3D distribution. The filamentary structure captured by the DMC is depicted in red on the central and bottom rows through the representation of its ascending 1-manifolds, after cancellation of the persistence pairs at a significance level of  $0-\sigma$  (left, no simplification),  $2-\sigma$  (center) and  $4-\sigma$  (right). The central left panel nicely illustrates the strong influence of Poisson noise, as without simplification, filaments are basically detected almost everywhere in the distribution. This is particularly obvious when zooming on what was a dark matter halo in the former 3D distribution: whereas one can identify by eye a few obvious filaments connecting to the central clump, the algorithm (correctly) detects a swarm of local peaks and filaments locally created by random fluctuations in the distribution.

It is quite striking though how much applying the above described persistence based simplification procedure succeeds at selecting what one would intuitively define as a filament. Already, at a  $2-\sigma$  level (central and middle bottom frame), it is clear that the large scale network of filaments is correctly identified as well as the valley resulting from the projected cosmic voids of the non projected distribution (the ascending 2-manifolds associated to the minima, symbolized by the white disks). At a level of  $2-\sigma$ , the probability that a topological feature such as an arc in the DMC is the result of Poisson noise is  $\sim 5\%$ . At  $4-\sigma$ , this probability goes down to  $\sim 0.006\%$  and any arc in the DMC can therefore safely be considered a feature of the underlying distribution. The lower right panel shows the arcs of the DMC that link maxima (purple crosses) and saddle points (green triangles) at a significance





**Figure 13.** The filaments measured in a 2D distribution obtained by projecting the particles from a slice of an N-Body cosmological simulation. The initial discrete distribution, its Delaunay tessellation and a zoom on a Halo (from the upper central part of the distribution) are displayed on the top row, with colour corresponding to the DTFE density. The filamentary structure is traced in red on the middle row, as the geometry of the arcs remaining after cancellation of persistence pairs with significance less than  $0-\sigma$  (middle left),  $2-\sigma$  (center) and  $4-\sigma$  (middle right). A zoom around a projected Halo is shown on the bottom row. The white disks, green triangles and purple crosses stand for the minima, saddle-points and maxima respectively (notes that they only represented on some panels for clarity).

level of  $4-\sigma$  around the projection of a large dark matter halo. At that level, the intricate initial network is reduced to a very neat set of filaments branching on a central clump. Note that while the network itself is simplified, the resolution is preserved which for instance allows for the correct identification of the merger of two relatively noisy filaments on the top right corner while preserving a very

clean network on large scale (central right frame).

The application to 3D distribution and in particular galaxy catalogues and large scale N-body simulation is presented in the companion paper, Sousbie, Pichon, Kawahara (2010).

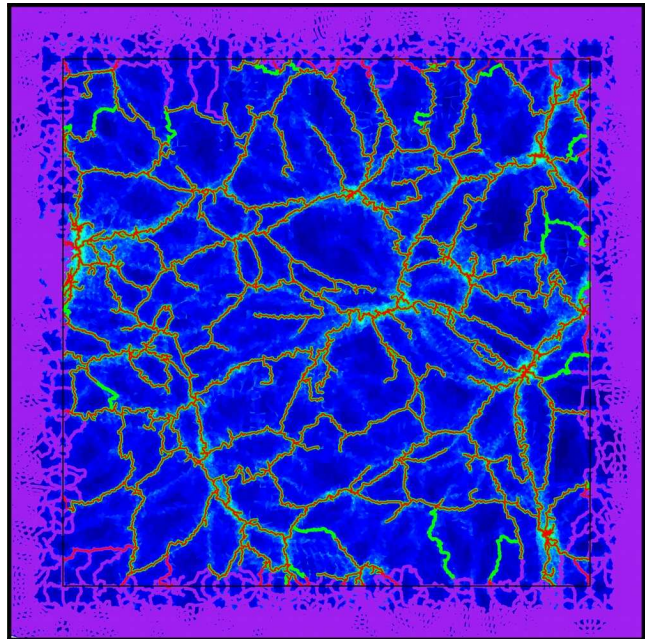


## 7 BOUNDARY CONDITIONS AND TECHNICALITIES

### 7.1 Boundary conditions

Whereas boundary conditions are not a concern in Morse theory, as it is defined over infinite or borderless spaces, things are clearly different when one tries applying it to real data sets. The easiest case corresponds to periodic boundary conditions data, such as those encountered in N-body simulation of matter distribution on large scales in the Universe. Because it is impossible to simulate the whole Universe and gravity has an infinite outreach, periodic boundary conditions are often used as a trick to obtain a smooth gravitational potential and emulate the isotropy of space within a restricted volume usually shaped as a box. Enforcing periodic boundary conditions over a cube basically amounts to assimilating opposite faces, any object leaving the cube through one face immediately entering the opposite one. Mathematically speaking such a space is called a torus  $\mathbb{T}^d$ , where  $d$  is the number of dimensions, and Morse theory readily applies to such spaces. From a practical point of view, we use the periodic exact 3D periodic boundary conditions Delaunay tessellation (Caroli & Teillaud 2010) implemented in CGAL<sup>12</sup> when the distribution is three dimensional. We also implemented our own periodic boundary conditions within CGAL for  $d \neq 3$  case using a simpler, though less rigorous and optimized, technique. This method basically consists in building a larger distribution by replicating a fraction of the box to extend each boundary, computing the delaunay tessellation over this extended domain and then identifying the identical  $k$ -faces crossing opposite faces of the initial box (the Delaunay tessellation of figure 13 was obtained using such method).

Of course, periodic boundary conditions only apply to periodic data sets, which is usually not the case of observational data, and one therefore needs to treat the boundaries of the distribution with special care. The simplest way to do so consists in transforming the definition domain of the data set into a boundaryless domain, a procedure called compactification. Usual compactification techniques consist in transforming the definition domain into a sphere by adding a point at infinity and attaching it to each boundary cell of the delaunay tessellation or transforming it into a torus, practically making it periodic by replicating a mirror image of the distribution through its boundaries. Both of these methods have pros and cons. Whereas sphere compactification is easy to build, whatever the geometry of the initial data set, it tends to pollute measurements around the boundaries by affecting the discrete gradient computation, therefore creating numerous fake manifolds and arcs that have to be ignored. This is not the case of the torus compactification, which creates relatively smooth conditions close to the boundaries, but it may only be easily implemented on cubic boxes and requires replicating the data set a large number of times (27 times in 3D), significantly increasing



**Figure 14.** Illustration of the computation of the DMC with non periodic boundary condition. The boundary of the initial 2D distribution are delimited by the thin black square, and any particle in the distribution within a distance of 10% the initial domain size is mirrored (see also figure 13). The thick green and purple network shows the  $3\text{-}\sigma$  filaments measured in the non-periodic distribution, the purple part being discarded after topological simplification as belonging to the boundary. The filaments obtained in the periodic boundary situation are displayed in red for comparison.

the necessary computational time and resources accordingly.

Our implementation of the boundary conditions is a hybrid between the sphere and torus compactification that tries to preserve the advantages of both while getting rid of their limitations. The idea is that the torus compactification is efficient because of the relatively natural extrapolation of the density field it allows, which as a result does not affect the computation of the discrete gradient at large distance from the boundary. We therefore allow the user to choose what fraction of the distribution should be mirrored on each boundary (a value around 10 ~ 15% of the initial distribution size seems to work fine), and then apply a sphere compactification on the new distribution by adding a point at infinity, with minus infinite density, that forms simplexes with the new boundary of the enlarged distribution. We then proceed by tagging as “boundary” any  $k$ -simplex of the Delaunay tessellation that contains a replicated vertex, the infinite vertex, or whose DTFE density may have been affected by the distribution outside the definition domain. This last condition in fact prevents the boundary simplexes, whose DTFE density may be wrong, to affect the resulting DMC and we determine which simplexes may be affected by checking whether the circumsphere of each highest dimensional  $d$ -simplex intersects the boundary, in which case it is, with all its faces and vertices, tagged as “boundary”.

<sup>12</sup> CGAL is the C++ Computational Geometry Algorithm Library, see <http://www.cgal.org>



Note that one has to pay particular attention to the boundaries during the topological simplification process as the persistence of critical pairs formed with a boundary simplex have spurious persistence ratios. The point at infinity is special, it has a minus infinite density, and is allowed to form persistence pairs with as many vertices as necessary, any of those persistence pairs having infinite persistence. The persistence pairs formed between the non-infinite boundary simplexes and those within the valid part of the distribution are treated normally during the simplification process, but the topological feature they form are nevertheless spurious and any persistence pair with at least one critical simplex on the boundary is therefore deleted<sup>13</sup> after topological simplification. The whole process is illustrated on figure 14 in the 2D case, using the same distribution as that of figure 13. On that picture, the filamentary structure detected at  $3-\sigma$  is represented for the same distribution when it is considered periodic (thin red network) and non-periodic (thick green network), and it is clear that both mostly agree. One can see nonetheless that, as should be expected, the small portions of the (red) periodic network crossing the boundaries cannot be detected in the non-periodic case and that a few portions of (green) filaments lying slightly farther away are detected only in the non-periodic case. This results from the fact that persistence pairs of distant critical simplexes may be different in the non-periodic distribution because the  $k$ -cycles coupling them are not allowed to cross the boundaries. As a result, the persistence ratios of certain critical points may differ in both cases, and they may therefore still exist at  $3-\sigma$  level in the non-periodic case while they were canceled at  $2.5-\sigma$  level in the periodic one.

## 7.2 Smoothing the manifolds

Because the scale resolution of practical samples is always limited, so is the resolution of the ascending and descending manifolds of the DMC. When identifying the filaments, voids or walls in cosmological distributions, their precise shape therefore becomes arbitrary at scales lower than the initial sampling resolution. Within our implementation, the DMC features are sub-sets of the initial Delaunay tessellation or of its dual Voronoi tessellation and the identified structures therefore naturally tend to adapt to the measured sample much better than they would if one was using a regular sampling grid for instance. The DMC is nevertheless always computed at the sampling resolution limit, and its geometry is mainly dictated by Poisson noise on that scale. It may therefore be desirable to have a way to enforce some continuity and differentiability even at the price of a loss of resolution (*e.g.* for instance for representation purposes). The smoothing method that we use is pretty much the same as that presented in Sousbie et al. (2009) as it presents the advantages of being simple, robust and fast. The idea involves smoothing the filaments individually by fixing the critical points and averaging the position of each non-fixed segment's endpoint with the position of its closest neighbouring endpoints a given number of times. Within our im-

plementation, a filament is defined as a sequence of  $N$  linked vertexes, the vertexes corresponding to the centers of mass of simplexes in the Delaunay tessellation. Let  $x_j^i$  be the  $j^{\text{th}}$  coordinate of  $i^{\text{th}}$  vertex. Then, after smoothing, its new coordinates  $y_j^i$ , are computed as:

$$y_k^i = A^{ij} x_k^j, \quad (10)$$

with

$$A^{ij} = \begin{cases} 3/4 & \text{if } i = j = 0 \text{ or } i = j = N, \\ 1/2 & \text{if } i = j, \\ 1/4 & \text{if } i = j + 1 \text{ or } i = j - 1, \\ 0 & \text{elsewhere,} \end{cases} \quad (11)$$

where equation (10) is applied  $s$  times in order to smooth over  $s$  simplexes in the simplicial complex. The corresponding smoothing length is naturally adaptive and such a smoothing ensures continuity of the filaments location over  $s$  Delaunay simplexes. Note that it is very easy to adapt this method to the ascending and descending manifolds of the DMC (*i.e.* the voids, walls, ...) as any of them is defined as a simplicial complex within our implementation. The position of each vertex in a manifold can therefore similarly be averaged with that of its neighbors  $s$  times to obtain sufficient smoothness.

## 7.3 Essential implementation issues

Finally, we close this section by presenting two technical issues that are essential to a practical implementation of the algorithm.

### 7.3.1 Cancellation order

When canceling persistence pairs, the order in which the pairs are canceled has a crucial importance, both in terms of computational time and memory consumption. This is especially true in 3D. In fact, whereas the number of arcs linking a given 1-saddle (resp. 2-saddle) and a maximum (resp. minimum) is always 2, there is no bound on the number of arcs between two saddle points of different types. Following the arc redirection algorithm described in section 6.2, the cancellation of two saddle-points of different type may therefore lead to a dramatic increase in the total number of arcs in the complex. Let  $P$  and  $Q$  be the 1-saddle and 2-saddle respectively. Then  $P$  is linked to  $P_\uparrow = 2$  maxima and  $P_\downarrow$  2-saddles, while  $Q$  is linked to  $P_\uparrow$  1-saddles and  $P_\downarrow = 2$  minima. The cancellation therefore creates  $N_C = (P_\downarrow - 1)(Q_\uparrow - 1)$  arcs and destroys  $N_d = 2 + 2 + P_\downarrow + Q_\uparrow - 1$  arcs, and the number of additional arcs after cancellation is  $N = N_C - N_d \propto P_\downarrow Q_\uparrow$  for large values of  $P_\downarrow$  and  $Q_\uparrow$ . This means that the number of arcs in the complex may temporarily increase quadratically, and it is not uncommon to obtain saddle points with hundreds of thousands of arcs at a given moment<sup>14</sup>. Within our implementation, we therefore always cancel the pair  $\{P, Q\}$ , with  $P$  the critical point of highest type,

<sup>13</sup> we really mean *deleted* in that case, and not canceled as a regular pair would be

<sup>14</sup> This does not mean that hundreds of thousands of arcs will be present in the simplified complex though, as a single maximum/1-saddle or a minimum/2-saddle may later cancel all those arcs leading to a dramatic decrease in the total number.

that minimize the number  $N$  of created arcs first, with  $N = N_c - N_d = (P_\downarrow - 1)(Q_\uparrow - 1) - P_\downarrow - P_\uparrow - Q_\downarrow - Q_\uparrow + 1$ .

### 7.3.2 Impossible cancellations

There exist special configurations where two critical points are linked by two or more different arcs (think for instance of the circular crest around the crater of a volcano). Those particular configurations cannot be canceled, as applying a discrete gradient reversal (see section 6.2) would result in the formation a V-path (*i.e.* discrete integral line) that loops onto itself; this is impossible as a V-path is a strictly decreasing alternating sequence of  $k$ -simplexes (see definition 3.7). This is not a problem for maximum/1-saddle and minimum/2-saddle cancellation though, as such persistence pairs cannot be formed if the critical points are linked by more than one arc (taking the example of the volcano, the highest point on its crest is a maximum, which are always positive (creating), and this is also the case of the lowest point on the crest which is a positive saddle point, as it creates the ring formed by the crest around the volcano). Yet the 3D case of a 1-saddle/2-saddle persistence pair is different, as nothing prevents such configurations to occur. In practice, such configuration do not arise naturally and we noticed that using the order for canceling pairs defined previously drastically reduces the number of occurrence of such non-cancellable configurations (of order  $\sim 10$  for a  $128^3$  particles simulation cut at  $4-\sigma$ ). For those few remaining pairs, we offer the possibility in our implementation to skip them or force their removal after only keeping one of the arcs between the critical points within the persistence pair. This last option is the preferred one, and although it seems difficult to justify from a theoretical point of view, the fact that the occurrence of non-cancellable pairs depends on the precise cancellation order suggests that it is acceptable to do so (note that the consequences on the resulting Morse-Smale complex are quite minimal anyway).

## 8 CONCLUSION

We presented a method that allows the scale-free and parameter-free coherent identification of all types of 3D astrophysical structure in potentially sparse discretely sampled density fields such as N-body simulations or observational galaxy catalogues. The method is based on Morse theory (section 2), *discrete* Morse theory (section 3) and persistence theory (section 4), and the implementation of the corresponding algorithm was detailed in sections 5, 6 and 7. In particular, our specific algorithm was designed with astrophysical applications in mind, as it directly applies to the delaunay tessellation of point set samples<sup>15</sup>, and we paid a particular attention to the computation of the discrete Morse function so that it correctly represents the

underlying DTFE density. From this discrete Morse function, DisPerSE basically computes the discrete Morse Smale complex of the density function and uses it to identify structures: the ascending 3, 2, 1 and 0 manifolds of the theory being identified to the voids, walls, filaments and clusters respectively. The implementation was designed so that each component of the cosmic web and its geometry can be easily identified and studied as individual objects or as group of objects and so that their relationship can be easily recovered: one can for instance identify the voids bordering a given wall or the clusters at the extremities of a given filament. Moreover, as the persistence criteria was re-casted in terms of confidence level with respect to noise, it make DisPerSE very easy to use, as it is the only parameter required to identify structure at optimal resolution. It shows a great deal of potential for astrophysical applications, for the following reasons that distinguish it from traditional methods:

- (i) It applies directly to discrete data sets via their Delaunay tessellation, which makes it scale free and allows the identified structures to always be defined down to the resolution limit of the sample.

- (ii) It is based on *discrete* Morse theory, which means that, as opposed to methods based on *smooth* Morse theory, the mathematical formalism does apply rigorously to the type of data sets one usually have to deal with in astrophysics. This implies that the well studied formalism of Morse theory readily applies to the numerically identified structures (which is not the case of watershed based methods for instance, see appendix A).

- (iii) All the different types of structures are defined coherently: triangulated space can basically be divided into sets of volumes, surfaces, curves and points that correspond to voids, walls, filaments and clusters respectively. Each structure is identified *individually*, and the cosmic web can therefore for instance be rigorously divided into individual filaments, each corresponding to a given saddle-point.

- (iv) It readily takes into account sampling and Poisson noise via persistence theory, allowing the user to define a detection confidence level in term of “number of sigmas” and provides the corresponding simplification of the DMC. As shown in Sousbie, Pichon, Kawahara (2010), this fact actually produces results obtained in highly sampled simulations and sparse galaxy catalogues which are qualitatively very similar, opening the way to a direct comparison of the properties of the cosmic web in simulations and observational catalogues.

- (v) Because the foundation of the method is based on topology and uses persistence theory, it also allows for a very robust computation of topological invariants such as Betti numbers or the Euler characteristic; this is possible even in the presence of an important shot noise, and without having to define any smoothing scale; it therefore takes into account the truly multi-scale nature of the cosmic web (see Sousbie, Pichon, Kawahara (2010)).

Application to 3D cosmic simulated and observed data sets are presented in the companion paper, Sousbie, Pichon, Kawahara (2010). Let us emphasize however that even if there is a wide range of application in astrophysics already, the domain of application of DisPerSE is undoubtedly wider than the cosmic large scale structures.

<sup>15</sup> in fact, the algorithm can also be used directly over structured regular meshes and we implemented a version that works directly on a regular grid.

## Acknowledgements

The author gratefully acknowledges support from JSPS (Japan Society for the Promotion of Science) Postdoctoral Fellowship for Foreign Researchers award P08324.

The author thanks C. Pichon for a careful reading and commenting of the manuscript, H. Kawahara for his fruitful comments and Y. Suto for his constant help and support.

This work was made possible through an extensive usage of the Yorick programming language by D. Munro (available at <http://yorick.sourceforge.net/>) and also CGAL, the Computational Geometry Algorithms Library, (<http://www.cgal.org>), which was used to compute the Delaunay tessellations.

## REFERENCES

- Abazajian K. N., et al., 2009, *ApJ Sup.*, 182, 543 3
- Aragon-Calvo M. A., Platen E., van de Weygaert R., Szalay A. S., 2008, ArXiv e-prints 2, 30, 31
- Aragon-Calvo M. A., van de Weygaert R., Araya-Melo P. A., Platen E., Szalay A. S., 2010, *MNRAS*, 404, L89 2, 30
- Aragón-Calvo M. A., van de Weygaert R., Jones B. J. T., 2010, *MNRAS*, pp 1270+ 2
- Aubert D., Pichon C., Colombi S., 2004, *MNRAS*, 352, 376 1
- Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, *ApJ*, 304, 15 2
- Bertschinger E., 1985, *ApJ Sup.*, 58, 1 1
- Beucher S., Lantujoul C., 1979, in Proceeding of International Workshop on Image Processing, Real-Time Edge and Motion Detection/Estimation Use of watersheds in contour detection. pp 17–21 30
- Bond J. R., Kofman L., Pogosyan D., 1996, *Nature*, 380, 603 1
- Caroli M., Teillaud M., 2010, in , CGAL User and Reference Manual, 3.6 edn, CGAL Editorial Board 26
- Cohen-Steiner D., Edelsbrunner H., Morozov D., 2006, in , Computational geometry (SCG'06). ACM, New York, pp 119–126 18
- Colberg J. M., Pearce F., Foster C., Platen E., Brunino R., Neyrinck M., Basilakos S., Fairall A., Feldman H., Gottlöber S., Hahn O., Hoyle F. e. a., 2008, *MNRAS*, 387, 933 30
- Colless M., Peterson B. A., Jackson C., Peacock J. A., Cole S., Norberg P., Baldry I. K., Baugh C. M., Bland-Hawthorn J., Bridges T., Cannon R. e. a., 2003, ArXiv Astrophysics e-prints 1
- de Lapparent V., Geller M. J., Huchra J. P., 1986, *ApJ Let.*, 302, L1 1
- Delfinado C. J. A., Edelsbrunner H., 1995, *Comput. Aided Geom. Design*, 12, 771 34
- Edelsbrunner H., Harer J., Natarajan V., Pascucci V., 2003, in SCG '03: Proceedings of the nineteenth annual symposium on Computational geometry Morse-smale complexes for piecewise linear 3-manifolds. ACM, New York, NY, USA, pp 361–370 30
- Edelsbrunner H., Harer J., Zomorodian A., 2003, *Discrete Comput. Geom.*, 30, 87 30
- Edelsbrunner H., Letscher D., Zomorodian A., 2000, in , 41st Annual Symposium on Foundations of Computer Science (Redondo Beach, CA, 2000). IEEE Comput. Soc. Press, Los Alamitos, CA, pp 454–463 2, 18
- Edelsbrunner H., Letscher D., Zomorodian A., 2002, *Discrete Comput. Geom.*, 28, 511 2, 10, 32, 34
- Forero-Romero J. E., Hoffman Y., Gottlöber S., Klypin A., Yepes G., 2009, *MNRAS*, 396, 1815 2
- Forman R., 1998a, *Math. Z.*, 228, 629 8
- Forman R., 1998b, *Adv. Math.*, 134, 90 2, 7
- Forman R., 2002, *Sém. Lothar. Combin.*, 48, Art. B48c, 35 pp. (electronic) 2, 7, 12, 23
- Gay C., Pichon C., Le Borgne D., Teyssier R., Sousbie T., Devriendt J., 2010, *MNRAS*, 404, 1801 2
- Gottloeber S., 1998, in V. Mueller, S. Gottloeber, J. P. Muecket, & J. Wambsganss ed., Large Scale Structure: Tracks and Traces Galaxy Tracers in Cosmological N-Body Simulations. pp 43–46 1
- Gyulassy A., 2008, PhD thesis, Univ. California Berkeley 2, 8, 12, 13
- Hahn O., Porciani C., Carollo C. M., Dekel A., 2007, *MNRAS*, 375, 489 2
- Hatcher A., 2002, *Algebraic topology*. Cambridge University Press, Cambridge 32
- Hoffman Y., Shaham J., 1982, *ApJ Let.*, 262, L23 1
- Huchra J. P., Geller M. J., 1982, *ApJ*, 257, 423 1, 10, 17
- Icke V., 1984, *MNRAS*, 206, 1P 1
- Jost J., 2008, *Riemannian geometry and geometric analysis*, fifth edn. Universitext, Springer-Verlag, Berlin 2
- Kirshner R. P., Oemler Jr. A., Schechter P. L., Sackett P. A., 1981, *ApJ Let.*, 248, L57 1
- Lewiner T., 2002, Master's thesis, Department of Mathematics, PUC-Rio 13
- Milnor J., 1963, *Morse theory*. Based on lecture notes by M. Spivak and R. Wells. Annals of Mathematics Studies, No. 51, Princeton University Press, Princeton, N.J. 2
- Neyrinck M. C., 2008, *MNRAS*, 386, 2101 2, 24
- Neyrinck M. C., Gnedin N. Y., Hamilton A. J. S., 2005, *MNRAS*, 356, 1222 1
- Novikov D., Colombi S., Doré O., 2006, *MNRAS*, 366, 1201 2
- Okabe A., ed. 2000, *Spatial tessellations : concepts and applications of voronoi diagrams* 7
- Platen E., van de Weygaert R., Jones B. J. T., 2007, *MNRAS*, 380, 551 2, 30
- Platen E., van de Weygaert R., Jones B. J. T., 2008, *MNRAS*, 387, 128 30
- Pogosyan D., Bond J. R., Kofman L., Wadsley J., 1996, in Bulletin of the American Astronomical Society Vol. 28 of Bulletin of the American Astronomical Society, The Cosmic Web and Filaments in Cluster Patches. pp 1289+ 1
- Pogosyan D., Pichon C., Gay C., Prunet S., Cardoso J. F., Sousbie T., Colombi S., 2009, *MNRAS*, 396, 635 31
- Robins V., 2000, PhD thesis, Department of Applied Mathematics, University of Colorado 10
- Roerdink J. B. T. M., Meijster A., 2000, *Fund. Inform.*, 41, 187 30
- Schaap W. E., van de Weygaert R., 2000, *A&A*, 363, L29 2, 7, 12, 24, 30
- Sousbie T., Colombi S., Pichon C., 2009, *MNRAS*, 393, 457 2, 3, 27, 30, 31
- Sousbie T., Pichon C., Colombi S., Novikov D., Pogosyan



- D., 2008, *MNRAS*, 383, 1655 2
- Sousbie T., Pichon C., Courtois H., Colombi S., Novikov D., 2008, *ApJ Let.*, 672, L1 2
- Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, *MNRAS*, 328, 726 1
- Stoica R. S., Martínez V. J., Mateu J., Saar E., 2005, *A&A*, 434, 423 2
- Stoica R. S., Martínez V. J., Saar E., 2010, *A&A*, 510, A38+ 2
- Tweed D., Devriendt J., Blaizot J., Colombi S., Slyz A., 2009, *A&A*, 506, 647 1
- Zomorodian A. J., 2009, *Topology for computing*. Vol. 16 of Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge 2, 18, 32

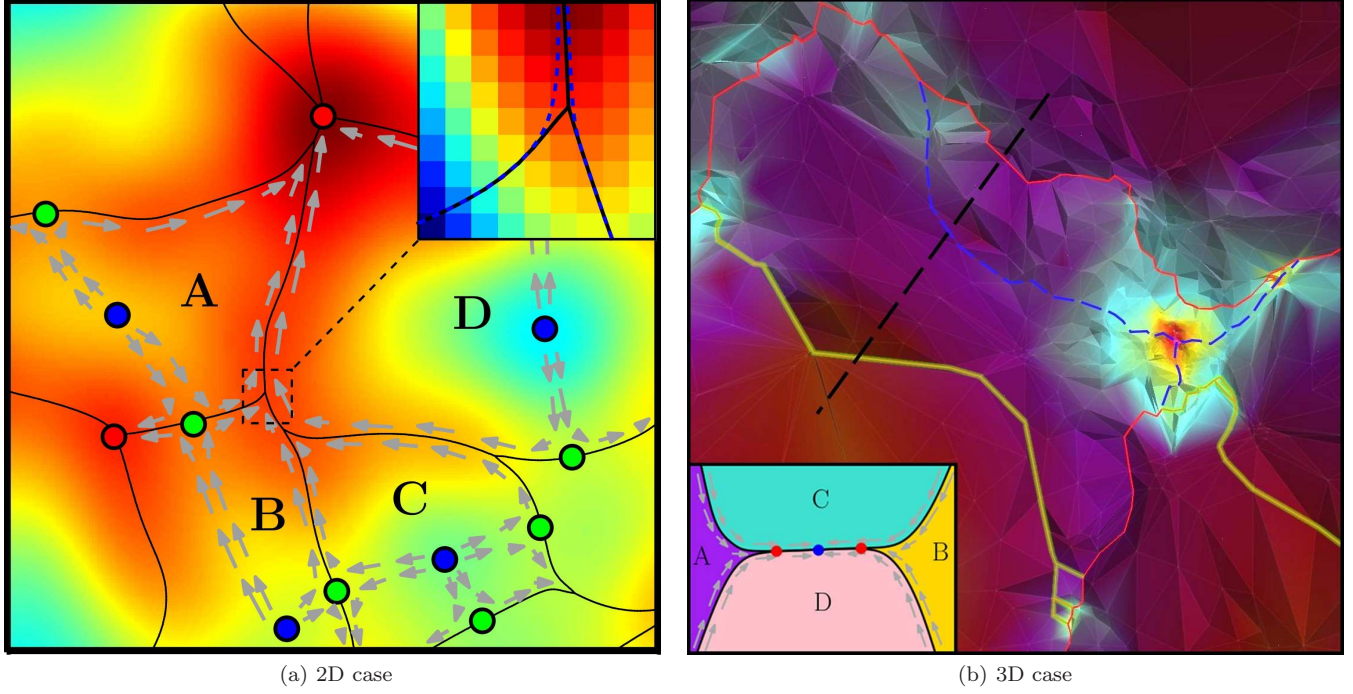
## APPENDIX A: APPLICABILITY OF MORSE THEORY TO PRACTICAL DATA-SETS

There exist a large number of methods to reconstruct a smooth density field from the discrete sample of galaxies in a catalogue or a dark matter particles distribution in a cosmological simulation. Whether one uses a simple constant resolution uniform grid to sample the original distribution or a more sophisticated scale free method such as DTFE (Schaap & van de Weygaert 2000), that is able to reconstruct the unbiased density field over the full dynamic range of the sample, the initial sampling always defines some lower scale resolution below which one is free to infer the behavior of the distribution. As the constraints undergone by a Morse function (definition 2.2) are essentially local (continuity, differentiability and non degeneracy of the critical points), one could imagine designing some sophisticated interpolation scheme that would enforce Morse properties on the distribution. In practice, designing such an interpolation scheme seems extremely difficult though and to our knowledge, this kind of solution have never been implemented. Another solution consists in relaxing Morse conditions by computing the manifolds and Morse complex of a non-Morse function, and later correct for this omission by enforcing the correct combinatorial properties on the pseudo Morse complex (see definition 2.8.1). The approach has been successfully developed by Edelsbrunner et al. (2003) and Edelsbrunner et al. (2003) in the 2D and 3D case respectively, but at the price of a very high algorithmic complexity. In fact, whereas the method for the 2D case has been implemented and tested, there exist no implementation to date in the case of a 3D function, although the method has been mathematically proved to be correct. Another more radical approach simply consists in abandoning the idea of rigorously computing the Morse-complex and rather rely directly on a pseudo-Morse complex. A pseudo-Morse complex is an approximation of a Morse complex and its combinatorial properties are not guaranteed by Morse theory anymore. This is mainly the result of a fundamental property of the paths defined by following the gradient arrows, the so-called integral lines, being violated: they are not guaranteed not to cross anymore, as they should with a Morse function (see definitions 2.3 and 2.3.1). The second approach recently became relatively popular in astrophysics as a way to identify

cosmologically significant structures, mainly using the Watershed transform. The Watershed technique (see Beucher & Lantujoul (1979); Roerdink & Meijster (2000)) was first applied to this kind of problem by Platen et al. (2007) as a mean of identifying voids in large scale structures (see also Platen et al. (2008), Colberg et al. (2008) or Aragon-Calvo et al. (2010)), it was latter extended to the identification of walls and filament through a pseudo Morse complex by Sousbie et al. (2009) and it is also used by Aragon-Calvo et al. (2008). But although promising, these techniques seem to be doomed by the lack of a consistent theory and therefore of a good understanding of the properties of the pseudo Morse complex, as illustrated in the following.

The watershed transform segments a field into isolated regions called basins, the analogs of the ascending manifolds of the minima (or equivalently 0-manifolds, see the top left frame of figure A1). The boundary of those basins delineate the walls (see bottom left frame) and the regions at the boundary of three basins describe the filaments as an approximation of the ascending manifolds of the first kind saddle points. We show on figure A1 how the fact that only a pseudo Morse complex is computed can lead to subtle but significant errors in the identification of the filaments in galaxy distribution. Figure 1(a) illustrates the problem in 2D, using a similar implementation as the one presented in Sousbie et al. (2009). A density field is sampled on a high resolution Cartesian grid and a watershed transform is applied, generating basins (labeled by letters). The filaments are therefore identified as the basins boundary (black curves) and form a pseudo Morse complex: a network that links the critical points together (the red, green and blue disks, standing for the maxima, saddle points, and minima respectively). One can see that the filaments seem correctly identified but according to Morse theory, if the Watershed transform yielded a correct Morse complex, field lines would only cross at critical points and the bifurcation points, located at the intersection of at least three basins (for instance *A*, *B* and *D* or *B*, *C* and *D*), would therefore be maxima. This is not the case on figure 1(a) because the function is not a Morse function, and its gradient lines may therefore intersect where the filaments seem to bifurcate (the gradient direction along critical lines is represented by the gray arrows). If the function complied to Morse criteria, these bifurcation points would actually look like the blue dashed line in the framed zoom in the upper right corner of the picture. This is not a significant problem for the identification of filaments in 2D, as it could theoretically be corrected for through some post-treatment, but as shown on figure 1(b), the consequences are more dramatic in the 3D case.

In order to assess the extent of this problem, a 3D multi-scale version of the probabilistic watershed transform presented in Sousbie et al. (2009) was implemented directly over a Delaunay tessellation computed from a discrete point sample. Each vertex of the tessellation is attributed a density using the DTFE method (Schaap & van de Weygaert 2000), and the probabilistic watershed transform is applied, using the natural neighborhood defined by the dual Voronoi cells to propagate the probabilities. Basically, the minima and maxima are identified as those vertice with



**Figure A1.** Illustration of a problem in the identification of the filaments when using the watershed technique to recover the Morse complex directly from a non Morse function. Figure 1(a): filaments (black curves) of a 2D field sampled at discrete locations over a high resolution grid, with gray arrows showing the gradient direction along those filaments. The maxima/saddle points/minima are represented as red/green/blue disks respectively and the letters designate regions delimited by filaments. Figure 1(b): filaments identified on a 3D delaunay tessellation of  $50h^{-1}$  Mpc large dark matter simulation with density computed using DTFE. The surface represents the boundary of a void, shaded according to the logarithm of the density and viewed from its corresponding minimum. The red and yellow curves show the filaments detected by a multi-scale watershed method. See main text for more explanations.

only higher or lower density neighbors respectively, and the probability that each vertex belongs to the integral line of a given extremum is computed according to Sousbie et al. (2009). This defines the watershed basins attached to minima and maxima (*i.e.* the void patches and peak patches according to the terminology of Sousbie et al. (2009), or the pseudo - ascending and descending 3-manifolds, according to Morse theory terminology). Figure 1(b) shows the triangulated interface between void patches (*i.e.* the boundary of the cosmological voids), computed over the delaunay tessellation of a sub-sampled  $512^3$  particles dark-matter cosmological simulation in a  $50h^{-1}$  Mpc box. This surface represents the density “walls” of the cosmic web, shaded according to the locally interpolated density. The surface is seen from the point of view of the minimum inside the void patch and one can identify a dark matter halo on the central-right part of the image. Following Sousbie et al. (2009) (see also Aragon-Calvo et al. (2008)) the filaments are identified as those segments located at the one dimensional interface of at least three different void patches, and are represented by the non-dashed red and yellow lines. It is clear on this picture that the yellow shaded lines are spurious as they do not correspond to any filament visible in the overdensity field projected onto the surface. One can also remark that the network does not pass through the local maximum located at the center of the halo, which should obviously be the case for a cosmological filament. Actually, a more reasonable network could be obtained by displacing the red lines to match the

blue dashed ones and removing the yellow shaded spurious identifications. The cause of those errors is actually similar to the one described in the previous paragraph for the 2D case: the density function does not comply to Morse criterion and its field lines may therefore cross. The sketch in the lower left illustrates what happens along the dashed black line, in the plane perpendicular to the surface: the void patches A and B are sandwiched between C and D, resulting in the identification of critical lines at the spurious intersection of ADC and BCD, symbolized by two red dots on the sketch, and the intersection of the dashed black line and the red and yellow critical lines on the 3D image. Actually, the only real critical line is at the true intersection of the four patches, symbolized by the blue dot on the sketch and blue dashed line on the picture (*i.e.* where the field lines really end, as represented by the gray arrows).

This tendency of the void and peak patches to get sandwiched between each other is perfectly natural and understood in Morse theory, and it is not a simple consequence of the particular selected sampling method, but rather of the fact that sampling is used at all. Moreover, it seems to be particularly the case in the large scale cosmological dark matter density fields, probably as a consequence of the nature of the initial Gaussian random field from which tiny perturbations evolve to form the cosmic web (see the discussion on bifurcations points in Pogosyan et al. (2009)). In short, this shows that the simple approach that consists in requiring filaments to be at the intersection of walls

which are at the intersection of voids is a bit naive as in practice, when the field is sampled and/or noisy, these boundaries do not have the right properties and do not trace the cosmic network correctly. These problems, among others, severely limit the domain of application of watershed based method (for instance it renders practically impossible their usage to count the number of filaments attached to a given halo, or the measurement of the physical properties of individual filaments) and demonstrate the necessity to adopt a different, mathematically more consistent approach.

## APPENDIX B: SIMPLICIAL HOMOLOGY

Homology theory studies the topological properties of a spaces (intuitively, its number of component, how they are connected or if holes exist ...). Roughly speaking, it does so by studying the properties of deformable chains and loops over these spaces and giving a method to relate them to sequences of Abelian groups, the so-called Homology groups. The goal of this section is only to give the reader enough intuitive understanding of its restriction to simplicial complexes - the weaker simplicial homology - to grasp the concept of topological persistence as introduced by Edelsbrunner et al. (2002). For that reason, although we give a few necessary mathematical definitions, we always try to explain them in a less formal and more intuitive manner. One could always refer to (Zomorodian 2009, chap. 4) for a very interesting and somewhat more rigorous introduction or Hatcher (2002) for a thorough reference.

In order to understand simplicial homology, one should first define the  $k$ -chain group over a simplicial complex  $K$  that contains  $p$  simplexes.

**Definition B.1. ( $k$ -chain group)** Let  $k \in \{0, \dots, d\}$  the dimension of the  $k$ -chain, then  $\{\sigma_1, \dots, \sigma_p\}$  is the set of all the  $k$ -simplexes in  $K$ . Any  $k$ -chain  $c_k$  can be written:

$$c_k = \sum_{i=1}^p n_i \sigma_i, \quad n_i \in \mathbb{Z}/2\mathbb{Z} = \{0, 1\}.$$

The  $k$ -chain group,  $C_k(K)$ , is the group with element  $c_k$  and addition defined as

$$c_k + c'_k = \sum_{i=1}^p (n_i + n'_i) \sigma_i.$$

In other words, a  $k$ -chain is a subset of the simplexes in  $K$  with dimension  $k$ . For a 3D simplicial complex such as the delaunay tessellation of a galaxy catalogue, it would be a set of vertice, segments, facets or tetrahedrons. Note that in this definition, although the more general case could be considered, the coefficients  $n_i$  are chosen to be positive integers modulo 2 which, as we will see, is sufficient to capture interesting topological properties. This means that a given simplex can only be absent or present once in a  $k$ -chain. Adding a simplex to a  $k$ -chain of  $C_k(K)$  that already contains it therefore results in its actual removal (the addition being performed modulo 2). This definition alone only relates simplexes of identical dimensions, but for different values of  $k$ , the  $C_k(K)$  are independent. The notion of topology (*i.e.* the connectivity of the simplexes in  $K$ ) can be

introduced through the definition of a boundary operator. Intuitively, the boundary of a simplex is the set of its faces:

**Definition B.2. (boundary operator)** Let  $v_i$  be  $k+1$  vertice of  $K$ , and  $\sigma = [v_0, v_1, \dots, v_k] \in C_k(K)$  a  $k$ -simplex, then the boundary of  $\sigma$  is

$$\partial_k(\sigma) = \sum_{i=0}^k [v_0, \dots, \hat{v}_i, \dots, v_k],$$

where  $\hat{v}_i$  means that vertex  $v_i$  is removed from the list. By extension, the boundary of a  $k$ -chain is defined as:

$$\begin{aligned} \partial : C_k(K) &\mapsto C_{k-1}(K) \\ c &\mapsto \partial c = \sum_{\sigma \in C_k(K)} \partial \sigma \end{aligned}$$

Following this definition, the boundary of a  $k$ -chain only contains the  $(k-1)$ -simplexes that are faces of exactly 1  $k$ -simplex in the chain. On figure B1 for instance, the segments in the orange contour (see upper right corner of the figure) are the boundary of the facets within the purple shaded area; all other purple shaded segments being faces of two facets, they cancel each other because of the addition modulo 2 in definition B.1. A very important property of the boundary operator is that  $\partial_{k-1}\partial_k = 0$ : the boundary of a boundary is void. This is intuitively easy to understand as a boundary is a cycle and cycles do not have boundaries. The orange boundary of figure B1 for instance forms a chain  $c_1$  that does not have boundary, as its segments all share the vertice at their extremity with exactly one other segment in  $c_k$ , and therefore appear twice when applying  $\partial_1$  to  $c_1$ . The subgroup of  $C_k(K)$  formed by the chains which are the boundary of a chain in  $C_{k+1}(K)$  is called the image of  $\partial_{k+1}$ .

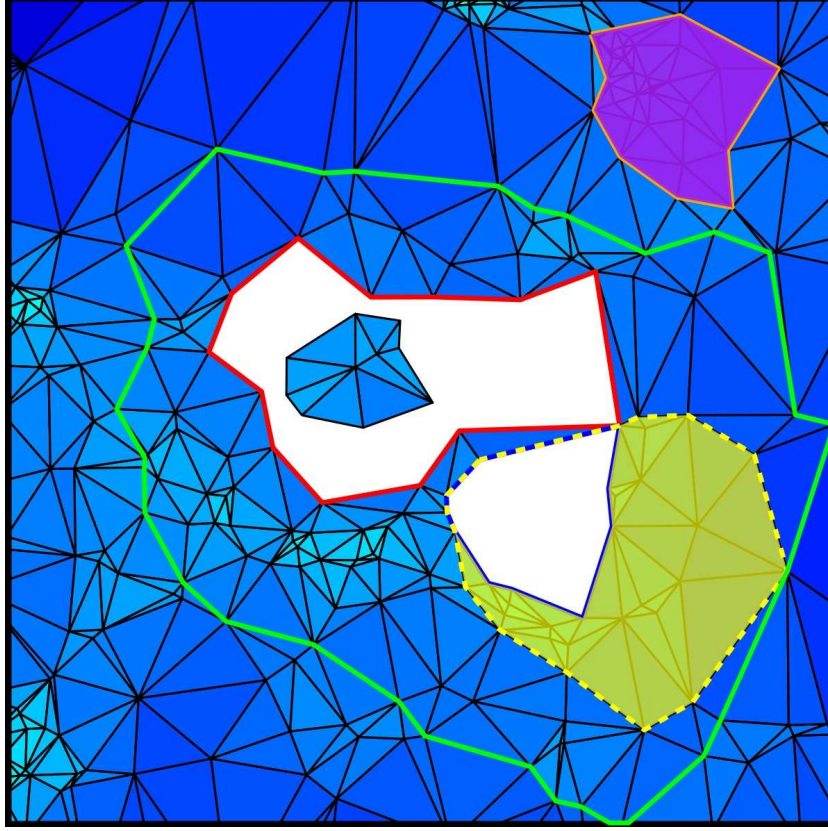
**Definition B.3. ( $k^{\text{th}}$  boundary group  $B_k$ )** Let  $B_k = \text{im } \partial_{k+1}$  be the image of  $C_{k+1}(K)$  under the boundary operator. Then  $B_k$  is a subgroup of  $C_k(K)$  called the  $k^{\text{th}}$  boundary group. Its elements form cycles called bounding cycles, and therefore do not have boundary.

On figure B1, a 1-chain of segments belongs to  $B_1$  if it is the boundary of a 2-chain of 2D facets, which is the case of the orange contour (boundary of the purple shaded facets) or the boundary of the yellow shaded area. This is nevertheless not the case of the green, red, blue and yellow dashed contours as no set of facets can have these contours as boundary due to the presence of the two holes. At best, the boundary of a 2-chain formed by a ring around a hole could include them, but it would necessarily contain additional cycles (the boundary of the hole). These contours are nevertheless cycles and therefore neither do they have boundaries. They all belong to the wider  $k^{\text{th}}$  cycle group:

**Definition B.4. ( $k^{\text{th}}$  cycle group  $Z_k$ )** Let  $Z_k = \ker \partial_k$  be the subset of  $C_k(K)$  whose image under  $\partial_k$  is the null  $(k-1)$ -chain. Then  $Z_k$  forms a subgroup of  $C_k$  called the  $k^{\text{th}}$  cycle group, and the  $k^{\text{th}}$  boundary group  $B_k$  is included in  $Z_k$ .

The elements of  $Z_k$  are any chain that form a cycle (or equivalently that have no boundary), and the green, red, blue and yellow dashed contours of figure B1 do belong to  $Z_1$ .





**Figure B1.** Illustration of 1-boundaries and 1-cycles of a 2D simplicial complex extracted from a filtration of a Delaunay tessellation. The facets present in the filtration are colored with different shades of blue, depending on the local density, and there are two holes at this stage (white parts). See main text for explanations..

These elements are enough to get an idea of how simplicial homology works. It involves trying to count how many different types of cycles it is possible to define for each dimension. To achieve this, one first needs to define what one means by “different types of cycles”, and to do so, homology define an equivalence relation over the  $k$ -chains:

**Definition B.5. (simplicial homology)** Two cycles  $c$  and  $c'$  in the  $k^{\text{th}}$  cycle group  $Z_k$  are said to be homologous if there exist a bounding cycle  $b \in B_k$  such that:

$$c + b = c'.$$

This equivalence relation can be used to define the class of equivalence of  $z \in Z_k$ ,  $[z]$ , which contains all the elements of  $Z_k$  that are homologous to  $z$  (i.e. all  $z' \in Z_k$  that can be written  $z + b = z'$  with  $b \in B_k$ ).

In a nutshell, definition B.5 formalizes, for simplicial complex, the intuitive idea that two cycles are equivalent if they can be continuously deformed into each other. This definition is at the core of regular Homology theory. For instance, the 1-chains represented by the blue and yellow dashed contours of figure B1 are homologous, as one can obtain the yellow one by adding the boundary of the yellow shaded 2-chain to the blue 1-chain. At the contrary, the red and yellow dashed 1-chains are clearly not homologous as it is impossible to find a chain that is both a boundary of a 2-chain and transform one into the other through addition. This impossibility clearly comes from the fact that there exist holes in

the simplicial complex, and homology shows that the presence of these hole directly affects the maximum number of non homologous cycles one can create. This link can be established through the so called  $k^{\text{th}}$  Homology group, which elements are the sets of homologous  $k$ -chains:

**Definition B.6. ( $k^{\text{th}}$  Homology group)** The  $k^{\text{th}}$  Homology group is the group which elements are the sets of homologous  $k$ -chains. It is defined as the quotient group of the  $k^{\text{th}}$  cycle group  $Z_k$  by the  $k^{\text{th}}$  boundary group  $B_k$ :

$$H_k = Z_k / B_k = \ker \partial_k / \text{im } \partial_{k+1}.$$

An element  $h$  of  $H_k$  is represented by the class of equivalence  $[z]$  of all chains homologous to  $z \in Z_k$ .

In other words, on figure B1, an element of  $H_1$  could be represented by the blue 1-chains around the smaller hole, as well as chains homologous to it such as the yellow dashed one. Another element is the red 1-chain and its homologous chains, and yet another one is the class of equivalence of the green contour. But there is something different with the green 1-chains: it may not be homologous to the blue and red ones, but it could be obtained by adding to cycles homologous to the red and blue ones respectively. This leads us to the definition of the Betti numbers, the mean by which homology describe the topology of a space:

**Definition B.7. ( $k^{\text{th}}$  Betti number)** the  $k^{\text{th}}$  Betti num-

ber  $\beta_k$  is the rank of the free<sup>16</sup> part of  $H_k$ :

$$\beta_k = \text{rank} H_k = \text{rank} Z_k - \text{rank} B_k,$$

To put it simply, the  $k^{\text{th}}$  Betti number really is the minimal number of  $k$ -cycles equivalence classes (*i.e.* sets of homologous  $k$ -cycles) that one needs to generate any possible cycle through homology. Betti numbers are interesting because they are characteristic of the topological properties of a given space, and in that sense allow quantifying and comparing the topologies of different spaces.

## APPENDIX C: PERSISTENCE AND BETTI NUMBERS IN A SIMPLICIAL COMPLEX

In order to explain the computation of persistence pair over a simplicial complex we use figure C1, a figure inspired from Edelsbrunner et al. (figure 3 2002). Although the reader can always refer to page 37 for an explanation of the terminology, it is advisable to read appendix B for a quick introduction to simplicial homology. The initial discovery of persistence was triggered by the design of a simple algorithm to compute the Betti numbers over a filtration of a simplicial complex, first presented in Delfinado & Edelsbrunner (1995). A filtration of a simplicial complex (definition 4.2) is a concept related to the one of sub-level set (definition 4.1). Basically, it consists in a set of sub-complexes which are given a particular order. Figure C1 shows the sub-complexes  $K^i$  in a filtration  $F$  of a simplicial complex  $K$ , the index  $i$  being represented in the bottom left part of each box. It is the counterpart of a sub-level set in the sense that the arrival order of each simplex in the filtration can be defined by a function that affects a value to each simplex, in which case each sub-complex  $K^i$  in the filtration can be defined as the set of simplexes with value higher or lower than a given threshold  $v_i$ . Note that the complex  $K$  is always the last to enter the filtration, and is therefore represented in box number 17. In this particular filtration, the simplexes of  $K$  enter one at a time (we skipped a few steps for the sake of conciseness, as symbolized by the gray hatched box). This does not have to be the case in general though, but because each sub-complex in the filtration is a simplicial complex, a particular simplex may never enter a filtration before any of its faces. In each frame, the newly entering simplex is colored in red or blue, and the two numbers following the index are the Betti numbers  $\beta^i = (\beta_0^i, \beta_1^i)$  of  $K^i$ . As detailed in appendix B,  $\beta_0$  represents the number of components in a complex (*i.e.* how many separated “islands” exist) while  $\beta_1$  is the number of holes or, equivalently, the number of independent non-homologous 1-cycles one can create in  $K^i$  in the more sophisticated language of homology.

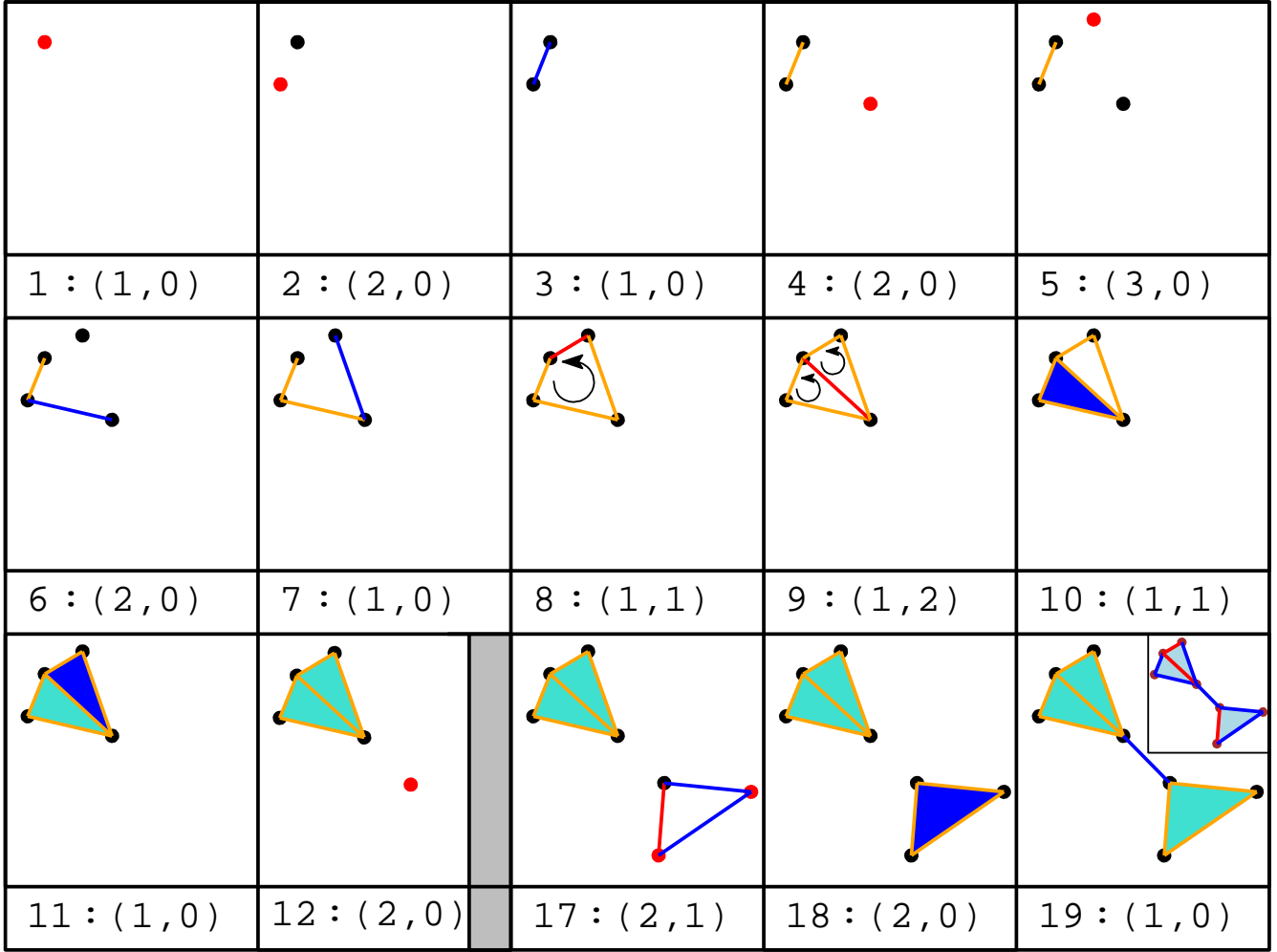
Let us see how Betti numbers can be computed using this particular algorithm.  $K^0$  is always the empty set, and so the algorithm starts with  $\beta^0 = (0, 0)$ . A vertex first enters  $F$  to form  $K^1$ , this adds one new component

in the filtration, but no one cycle can still be created, so  $\beta^1 = (1, 0)$ . As the entering vertex *created* a new component, it is represented in red and is labeled “positive”. Step 2 is essentially the same, and therefore  $\beta^2 = (2, 0)$ . In  $K^3$  though, the first segment enters  $F$ . Although we had two distinct components in  $K^2$ , the segment creates a link between them, and only one component remains. As one component was *destroyed*, the entering segment is represented in blue and labeled “negative”, and  $\beta_0$  decreases, leading to  $\beta^3 = (1, 0)$  again. Nothing special happens up to  $K^8$ , every entering vertex creating a new component therefore increasing  $\beta_1$  while each new segment destroys a components, therefore decreasing the value of  $\beta_0$ , leading to  $\beta^7 = (1, 0)$ . The segment entering  $K^8$  is different though, as it does not destroy any component: all the simplex in  $K^7$  were already linked and the new segment only links two vertices that already belonged to the same component. Actually, it *creates* a new class of 1-cycles (black rounded arrow) as it is now possible to draw a segments path that starts and ends at the same segment while passing through each other segment in the path only once (equivalently, it creates a hole within the cycle). The value of  $\beta_1$  is therefore increased and  $\beta^8 = (1, 1)$ . The entering segment is labeled “positive” and represented in red. The new segment in  $K^9$  is of the same kind: it creates a second hole, or equivalently a second class of cycles (black circular arrows) that is not homologous to the previous one. In fact, one cannot transform one into the other by adding the boundary of a set of facets, as there is no facet in the complex yet anyway. The entrance of a facet in  $K^{10}$  changes this fact, as this facet does fill one of the previously created hole: by adding the edges of this facet to the cycle created in  $K^8$  one obtains a cycle created in  $K^9$ , the two classes therefore becoming homologous (remember that by adding a simplex to a complex already containing it, one actually removes it). This leads to  $\beta_1$  being decreased and therefore  $\beta^{10} = (1, 1)$ . The filtration then goes on until all simplexes in  $K$  have entered, and  $\beta^{19} = (1, 0)$ .

Although we only presented a 2D example here, the procedure works for any number of dimensions and one can in general think of a  $k$ -cycle as the shell of a deformed  $(k+1)$ -dimensional sphere triangulated with  $k$ -simplexes, the simplest  $k$ -cycle being the faces of a  $(k+1)$ -simplex. The algorithm therefore consists in labeling each  $k$ -simplex of  $K$  as “positive” if it creates a  $k$ -cycle and “negative” if it destroys one when entering the filtration.<sup>17</sup> Going a bit farther, one can see that actually any cycle destroyed by an entering simplex were created earlier in the filtration. For instance, the segment that enters in  $K^3$  destroys the component created by the entering vertex in  $K^1$  or  $K^2$ . By convention, we will say that it destroys the most recently created, the vertex entering  $K^2$ . Identically, the new segment in  $K^6$  destroys the new component created in  $K^4$ , and the loop created in  $K^9$  is destroyed by the facet entering  $K^{10}$  while the facet entering

<sup>16</sup> The term “free” in the definition actually excludes some specific cycles that may exist when the space has torsion (think about a Möbius strip for instance)

<sup>17</sup> The question of how to decide whether a newly entered simplex actually belongs to a cycle or not is addressed in Edelsbrunner et al. (2002), but we do not present the method here as it is not essential to understand the concept of persistence. The implementation of such an algorithm is detailed in section 6.2.



**Figure C1.** Illustration of the filtration of a 2D simplicial complex. Each box represent one step of the filtration, with index  $i$  (lower left corner) and Betti numbers  $(\beta_0, \beta_1)$ . The gray hatched box symbolizes the fact that a few steps are not represented.

$K^{11}$  destroys the cycle created by the segment entering  $K^8$ . This defines pairs of negative and positive simplexes that create and destroy cycles, the partner of a positive (resp. negative)  $k$ -simplex being a negative  $(k-1)$ -simplex (resp. positive  $(k+1)$ -simplex). All the cycles can therefore be attributed some sort of “lifetime” in the filtration, equal to the index difference of their creating and destroying simplexes. This lifetime is called their persistence. In the case of figure C1 for example, the most persistent topological feature of  $K$  would be that  $K$  has two main components, joined by a central bridge: the segment entering  $K^{19}$  destroys the component created by the vertex entering  $K^{12}$ , the persistence of this topological feature therefore is  $19 - 12 = 7$ , which is larger than any other in the filtration. Of course, for a given complex, the persistence of each cycle (and actually the cycles themselves) depends on their precise order of arrival and what persistence really assesses is the topological properties of a function defined on the simplicial complex (*i.e.* the function that defines the order of arrival of the simplexes in the filtration).





## TERMINOLOGY

**Arc** An arc is a 1-cell: an integral line (or a V-path in the discrete theory) whose origin and destinations are critical points. The arcs of Morse-Smale complex comply to conditions 2.8.1, in particular, an arc always connects two critical points of order difference 1 (*i.e.* in 2D, a minimum and a saddle-point or a maximum and a saddle-point).

**$n$ -cell** A  $n$ -cell is a region of space of dimension  $n$  such that all the integral lines in the  $n$ -cell have a common origin and destination. The  $n$ -cells basically partition space into regions of uniform gradient flow (see definition 2.7)

**Coface** A coface of a  $k$ -simplex  $\alpha_k$  is any  $p$ -simplex  $\beta_p$ , with  $p \geq k$ , such that  $\alpha_k$  is a face of  $\beta_p$ . In 3D, the cofaces of a segment (*i.e.* a 1-simplex) are any triangle or tetrahedron (*i.e.* 2 or 3-simplex) whose set of summits (*i.e.* vertexes) contains the two vertexes at the extremities of the segment, as well as the segment itself. (see definition 3.2)

**Cofacet** A cofacet of a  $k$ -simplex  $\alpha_k$  is a coface  $\beta_{k+1}$  of  $\alpha_k$  with dimension  $k + 1$ . Equivalently,  $\alpha_k$  is a facet of  $\beta_{k+1}$ . (see definition 3.2)

**Critical point of order  $k$**  For a smooth function  $f$ , a critical point of order  $k$  is a point such that the gradient of  $f$  is null and the Hessian (matrix of second derivatives) has exactly  $k$  negative eigenvalues. in 2D, a minimum, saddle point and maximum are critical points of order 0, 1 and 3 respectively. (see definition 2.1)

**Critical  $k$ -simplex** A critical  $k$ -simplex is the equivalent in discrete Morse theory of the critical point of order  $k$  in its smooth counterpart. Note that in 2D, the equivalent of a minimum is a critical vertex (0-simplex), a saddle-point is a critical segment (1-simplex) and a maximum is a critical triangle (2-simplex). (see definition 3.5)

**Crystal** A crystal is a 3-cell: a 3D region delimited by 6 quads and 12 arcs, within which all the integral lines (or V-pathes in the discrete case) have identical origin and destinations.

**$k$ -cycle** A  $k$ -cycle in a simplicial complex corresponds to a  $k$  dimensionnal topological feature. in 3D, 0-cycles correspond to independant components, 1-cycles to loops and 2-cycles to shells (see definition 4.3 and appendix B)

**Discrete Gradient** A discrete gradient of a discrete Morse-Smale function  $f$  defined over a simplicial complex  $K$  pairs simplexes of  $K$  according to the rules of definition 3.6. Within a gradient pair, the simplex with lower value is called the tail and the other the head, and any unpaired simplex is critical (see definition 3.6).

**Discrete Morse-Smale complex (DMC)** The discrete Morse-Smale complex (DMC for short) is the equivalent of the Morse-Smale complex applied to simplicial complexes (see discrete Morse theory as introduced in section 3) (see definition 2.5).

**Discrete Morse-Smale function** A discrete Morse-Smale function  $f$  defined over a simplicial complex  $K$  associates a real value  $f(\sigma_k)$  to each simplex  $\sigma_k \in K$  and that obey the condition described in definition 3.4.

**Excursion set** see sub-level set.

**Face** A face of a  $k$ -simplex  $\alpha_k$  is any  $p$ -simplex  $\beta_p$  with  $p \leq k$ , such that all vertexes of  $\beta_p$  are also vertexes of  $\alpha_k$ . In 3D, the faces of a 3-simplex (*i.e.* a tetrahedron) are the tetrahedron itself, the 4 triangles that form its boundaries, the 6 segments that form its edges, and its 4 summits (*i.e.* vertexes). (see definition 3.2)

**Facet** A facet of a  $k$ -simplex  $\alpha_k$  is a face  $\beta_{k-1}$  of  $\alpha_k$  with dimension  $k - 1$ . The facets of a 3-simplex (*i.e.* a tetrahedron) are the 4 triangles (*i.e.* 2-simplexes) that form its boundaries (see definition 3.2)

**Filtration** A filtration of a simplicial complex  $K$  is a *growing* sequence of sub-complexes  $K_i$  of  $K$ , such that each  $K_i$  is also a simplicial complex. If the different  $K_i$  are defined by a discrete function  $F_\rho$  as the set of simplexes of  $K$  with value  $F_\rho(\sigma)$  less or equal to a given threshold, a filtration can be thought of as the discrete equivalent of a sequence of growing sub-level sets of a smooth function. (see definition 4.2)

**Gradient pair / arrow** A Gradient pair or arrow is a set of two simplex, one being the facet of the other, and such that they are paired within a discrete gradient. Within a gradient pair, the simplex with lower value is called the tail and the other the head.

**Integral line** An integral line of a scalar function  $\rho(\mathbf{x})$  is a curve whose tangent vector agrees with the gradient of  $\rho(\mathbf{x})$ . An integral line obeys properties 2.3 (see definition 2.3)

**Level set / Sub-level set** A level set, also called iso-contour, of a function  $\rho(\mathbf{x})$  at level  $\rho_0$  is the set of points such that  $\rho(\mathbf{x}) = \rho_0$ . The corresponding Sub-level set is the set of points such that  $\rho(\mathbf{x}) \geq \rho_0$  (see definition 4.1)

**Ascending/Descending  $p$ -manifold** Within a space of dimension  $d$ , an ascending  $p$ -manifold is the set of points from which, following minus the gradient, one reaches a given critical point of order  $d - p$ . A descending  $p$ -manifold is the set of points from which, following the gradient, one reaches a given critical point of order  $p$ . For instance, ascending 1-manifolds in 3D can be associated to the filaments, and ascending 3-manifolds describe the voids (see definition 2.4)

**Morse function** A Morse function is a continuous, twice differentiable smooth function whose critical points are non degenerate. In particular the eigenvalues of the Hessian matrix (*i.e.* the matrix of the second derivatives) must be non-null (see definition 2.2)

**Morse complex** The Morse complex of a Morse function is the set of its ascending (or descending) manifolds (see definition 2.5)

**Morse-Smale function** A Morse-Smale function is a Morse function whose ascending and descending manifolds intersect *transversely*. This means that there exist no point where an ascending and a descending manifold may be tangent (see definition 2.6 or 3.8 for the discrete case)

**Morse-Smale complex** The Morse-Smale complex is the intersection of the ascending and descending manifolds of a Morse-Smale function. One can think of the Morse-Smale complex as a network of critical points connected by  $n$ -cells, defining a notion of hierarchy and neighborhood among them. In particular, the geometry of the arcs (*i.e.* 1-cells) is determined by the

critical integral lines (*i.e.* integral lines that join critical points) and the order of two critical points connected by an arc may only differ by 1.

**Peak/Void patch** In 3D, a peak patch is a descending 3-manifold (*i.e.* the region of space from which, following the gradient, one reaches a given maximum), and a void patch an ascending 3-manifold (*i.e.* the region of space from which, following minus the gradient, one reaches a given minimum).

**Persistence** The persistence of a persistence pair (or equivalently of the corresponding  $k$ -cycle it creates and destroys) is defined as the difference between the value of the two critical points (or critical simplexes in the discrete case) in the pair. It basically represents its life time within the evolving sub-level sets, or filtration in the discrete case. (see section 4 and definition 4.4)

**Persistence pair** In the smooth context of a function  $\rho$ , persistence pairs critical points  $P_a$  and  $P_b$  of  $\rho$  that respectively create and destroy a topological feature (or  $k$ -cycle) in the sub-level sets of  $\rho$ , at levels  $\rho(P_a)$  and  $\rho(P_b)$ . In the discrete case of a simplicial complex  $K$ , a persistence pair is a pair of critical simplexes  $\sigma_a$  and  $\sigma_b$  of a given discrete function  $F_\rho(\sigma)$ , such that  $\sigma_a$  creates a  $k$ -cycle (*i.e.* topological feature) when it enters the filtration of  $K$  according to  $F_\rho$  and  $\sigma_b$  destroys it when it enters. (see section 4 or appendix C for more details)

**Persistence ratio** The persistence ratio of a persistence pair (or equivalently of the corresponding  $k$ -cycle it creates and destroys) is the ratio of the value of the two critical points (or critical simplexes in the discrete case) in the pair. Persistence ratio is preferred to regular persistence in the case of strictly positive functions such as the density field of matter on large scales in the universe. (see also the definition of persistence)

**Quad** A quad is a 2-cell : a 2D region delimited by four arcs within which all the integral lines (or V-pathes in the discrete case) have identical origin and destinations.

**$k$ -simplex** A  $k$ -simplex is basically the  $k$  dimensional analog of a triangle: the simplest geometrical object with  $k + 1$  summits, called vertex. It is the building block of simplicial complexes (see definition 3.1)

**Simplicial complex** A simplicial complex  $K$  is a set of simplexes such that if a  $k$ -simplex  $\alpha_k$  belongs to  $K$ , then all its faces also belong to  $K$ . Moreover, the intersection of two simplexes in  $K$  must be a simplex that also belongs to  $K$  (see definition 3.3)

**Vertex** A vertex is a 0-simplex or simply a point.

**V-path** A V-path is the discrete equivalent of an integral line: it is a set of simplexes linked by discrete gradient arrows and face/coface relation. Tracing a V-path basically consists in intuitively following the direction of the gradient pairs of a discrete gradient from a critical simplex to another. (see definition 3.7)